DOCUMENT RESUME

ED 094 358 CS 001 260

AUTHOR	Blanton, William E., Ed.; And Others
TITLE	Measuring Reading Performance.
INSTITUTION	International Reading Association, Newark, Del.
PUB DATE	74
NOTE	76p.
AVAILABLE FROM	International Reading Association, 800 Barksdale Road, Newark, Delaware 19711 (Stock No. 718, $3.50 non-member, $2.50 member)
EDRS PRICE	MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS	*Criterion Referenced Tests; Disadvantaged Youth; Elementary Education; Evaluation Criteria; Measurement Instruments; *Performance Contracts; *Reading Ability; *Reading Tests; *Test Selection

ABSTRACT

Designed to provide solutions to some of the problems related to measuring reading behavior, this publication explores some of the problems of test selection and usage which confront educators. Contents include "Reading Testing for Reading Evaluation" by Walter R. Hill, "Reading Tests and the Disadvantaged" by Thomas J. Fitzgibbon, "What Is Criterion-Referenced Measurement?" by Frank B. Womer, "Criterion-Referenced Tests: A Critique" by Frederick B. Davis, "Reading Tests and Performance Contracting" by Thomas P. Hogan, and "Comments on Impertinent Pertinent Problems with Content Validity in Reading Tests: A Response to Hogan" by J. Jaap Tuinman. (RB)

ED 094358

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

measuring reading performance

edited by

William E. Blanton
Roger Farr
J. Jaap Tuinman

Indiana University

INTERNATIONAL READING ASSOCIATION
800 Barksdale Road Newark, Delaware 19711

CS 001 260

INTERNATIONAL READING ASSOCIATION

OFFICERS
1974-1975

President Constance M. McCullough, California State University, San Francisco, California

Vice-President Thomas C. Barrett, University of Wisconsin, Madison, Wisconsin

Vice-President Elect Walter H. MacGinitie, Teachers College, Columbia University, New York, New York

Past President Millard H. Black, Los Angeles Unified School District, Los Angeles, California

Executive Director Ralph C. Staiger, International Reading Association, Newark, Delaware

DIRECTORS

Term expiring Spring 1975

Harold L. Herber, Syracuse University, Syracuse, New York
Helen K. Smith, University of Miami, Coral Gables, Florida
Grace S. Walby, Child Guidance Clinic of Greater Winnipeg, Winnipeg, Manitoba

Term expiring Spring 1976

Ira E. Aaron, University of Georgia, Athens, Georgia
Lynette Saine Gaines, University of South Alabama, Mobile, Alabama
Tracy F. Tyler, Jr., Robbinsdale Area Schools, Robbinsdale, Minnesota

Term expiring Spring 1977

Roger Farr, Indiana University, Bloomington, Indiana
Grayce A. Ransom, University of Southern California, Los Angeles, California
Harry W. Sartain, University of Pittsburgh, Pittsburgh, Pennsylvania

Copyright 1974 by the
International Reading Association, Inc.

Library of Congress Cataloging in Publication Data
Main entry under title:
Measuring reading performance.
Papers originally presented at a conference held by the IRA Test Committee at Indiana University in the fall of 1971.
Includes bibliographies.
1. Reading—Ability testing—Addresses, essays, lectures. I. Blanton, William, ed. II. Farr, Roger C., ed. III. Tuinman, J. Jaap, ed.
LB1050.M4 428'.4'07 74-11048
ISBN 0-87207-718-7

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

International Reading Association

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

CONTENTS

Foreword *v*

Preface *vi*

1 Reading Testing for Reading Evaluation *Walter R. Hill*

15 Reading Tests and the Disadvantaged *Thomas J. Fitzgibbon*

34 What Is Criterion-Referenced Measurement? *Frank B. Womer*

44 Criterion-Referenced Tests: A Critique *Frederick B. Davis*

51 Reading Tests and Performance Contracting *Thomas P. Hogan*

66 Comments on Impertinent Pertinent Problems with Content Validity in Reading Tests: A Response to Hogan *J. Jaap Tuinman*

The International Reading Association attempts, through its publications, to provide a forum for a wide spectrum of opinion on reading. This policy permits divergent viewpoints without assuming the endorsement of the Association.

FOREWORD

The consideration of evaluation in reading or in any other area evokes more questions than answers. Frequently the questioning is that of the educator in reaction to statements made by individuals whose prime qualifications lie in their having been evaluated—tested—at some more or less remote time, or in having gained some elective office.

One wonders how Ayres, Thorndike, Courtis, or other pioneers in the field of mental measurement would have reacted to the varied uses made of evaluative devices today. What would have been their reaction to the allocation of funding primarily based on pupil growth as measured by standardized reading tests? How would they have felt about the withdrawal of special funding for the least advantaged pupils in a community if they failed to achieve an arbitrarily determined standard?

Through what process does an educator or a member of a local school board equate the obligation to pay for services rendered with minimum pupil growth in a skill?

One last question: Do we no longer perceive the measurement of growth in a subject area as a tool to be used by teachers, pupils, administrators, and parents in determining how best to achieve the fundamental objective of the school—the most efficient learning by each pupil in the group?

This publication is designed to explore some of the problems of test selection and usage which confront educators today. The Association expresses its gratitude to the authors who have contributed to this volume in order that the goals of education may be better served through evaluation programs.

Millard H. Black, *President*
International Reading Association
1973-1974

PREFACE

The concern over the complaint, "Why can't Johnny read?" has reached an apex. Today, as never before, federal and state agencies, educators, and parents are asking, "What is the nature and extent of the reading problem?" "Who is responsible for it?" "What steps can be taken to remedy the problem?" and "How can we insure that the problem is corrected?" Attempts to answer these questions have led to the introduction of new terms such as "educational accountability," "performance contracting," "criterion-referenced tests," and "testing and the disadvantaged child" into the educational jargon of the reading specialist.

The IRA Test Committee, being concerned with sponsoring provocative discussions of new trends and practices affecting the measurement of reading, held a conference during the Fall of 1971 at Indiana University. The purpose of the conference was to consider some of the issues and problems behind the labels mentioned above. In developing the conference program, it seemed appropriate to invite a group of speakers who had particular expertise and experience in educational measurement to argue issues on the above topics that are still unresolved.

The papers and discussions presented herein speak for themselves and the reader will find that they constitute real contributions toward partial solutions to some of the problems related to measuring reading behavior within the contexts of performance contracting, criterion-referenced tests, and testing the disadvantaged child. We wish to acknowledge the cooperation shown by Indiana University and all speakers and discussants, as well as the continued interest of the members of the IRA Evaluation of Tests Committee.

WEB
RCF
JJT

READING TESTING FOR READING EVALUATION

Walter R. Hill
State University of New York at Buffalo

In reading, as in other areas of the educational program, there is an unfortunate, though human, tendency to overplay the role of testing while overlooking the broader essential functions of evaluation. Reading testing does perform a vital function in reading education—but as an integral and component part of reading evaluation. An understanding of the contributive operations of reading testing requires an appreciation of the other components. It may be useful to review briefly some operational definitions of *testing, measurement, assessment,* and *evaluation* as they apply to reading. These terms and their underlying concepts too often are employed as interchangeable parts. Properly employed, each is a necessary part in the machinery of reading education. At the same time, these parts are not meaningfully separated; their interdependency is presumed in the design of the reading program.

A *reading test* is concerned with sampling reading or reading-related behavior. Regardless of whether the results of the test are to be standardized into comparative norms, the procedures of a test should be standardized. There are a number of variations in the appearance and operations of reading tests, but to meet the requirements of a "test," each should present a uniform task to all examinees as well as provide some consistent means of comparing or interpreting an examinee's responses. A good reading test is highly valid and reliable. If the reading test elicits responses (behavior samples) which are quite typical of an examinee's nontest performance on those reading behaviors the test intends to measure, we say it has validity. If the reading test elicits results which are consistent with itself, we say it is reliable (Farr, 1969). A reading test must be reliable in order to be considered valid.

Reading measurement is concerned with the quantification of reading behavior (Lennon, 1962). It is concerned with answering the

questions of how much or to what degree, that is, with ascribing numbers to reading responses. Since a reading test utilizes a uniform task-procedure situation, it lends itself to the quantification of results. However, reading measurement is a facilitative tool of reading evaluation. The interpretation of quantified results will require some sort of evaluative reference. The two most viable referents utilized in educational testing are *norm-referent* (which involves comparing an examinee's scores with a distribution of scores obtained by others, usually educational peers, taking the same test) and *criterion referent* (which involves comparing an examinee's result with some preestablished standards of performance, frequently on an absolute basis).

There are many ways and situations in which reading and reading-related behavior may be *assessed* or systematically observed, especially by the classroom teacher. Although the usefulness of such observations is improved through the utilization of controlled situations and the quantification of results, it is not always possible or practical to structure these observations rigidly enough to meet the criteria of "testing." Some examples of useful but nontest-type observations are those obtained from progress charts, checklists, library or book-use records, interviews, anecdotal records, and autobiographies. In its broadest sense, reading assessment includes any empirical data gathering procedure from which observations pertinent to reading behavior may be obtained. In the collective sense, the term may be used to refer to the composite of such procedures, both planned and emerging, test and nontest, obtained about individual or group reading behavior. The accuracy of nontest assessment procedures may be improved through structure of procedure and quantification of observations. It is ironic that though authorities and formal courses in educational measurement have stressed the technical understanding of standardized test making, classroom teachers have had to depend largely upon informal assessment for functional guidance of pupil learning. In general, the informal assessment procedures utilized by classroom teachers have been of poor quality, but this can be changed through training. The expansion of reading assessment beyond standardized testing extends the number, variety, and functional quality of observations of reading behavior.

Reading evaluation necessarily involves judgment—an appraisal of behavior in terms of "how good" it is or how well it satisfies a desired outcome. The judgment may be normative: how much better or more satisfactory one pupil's behavior is when compared to that of other pupils. Or it can be a criterion judgment: did the behavior demonstrate a mastery of a specific objective of performance? The term frequently carries an administrative connotation. The reading evaluation program is the total or umbrella operational structure which

includes all school activity to gather, process, interpret, and react to data about the reading program and its pupils. It includes making judgments about curriculum, instruction, and materials as well as the evaluation of pupil performance and progress.

Reading testing frequently is confused with *reading diagnosis.* Since reading diagnosis is an important function of reading instruction, it deserves some attention in this brief glossary of reading evaluation. Reading diagnosis could be considered as a microform of reading evaluation. Usually, it is based upon multiple assessment of pupil reading or reading-related behavior. Some of these are quantified; others verbally descriptive. Some are test-derived measures. Diagnosis requires that judgment must be made about the adequacy of present reading performance. If the level of reading performance or rate of growth is judged to be inadequate, which is usually determined by comparing present reader achievement with reader expectancy or potential, the reading diagnosis may be extended to include an analysis of reading sub-behavior. Many authorities believe diagnosis should include the identification of likely causal factors contributing to the present deficiency in performance. In any case, the usual goal of the diagnostic case study is the prescription of corrective and remedial procedures and program.

An aspect of reading diagnosis too often overlooked is that it is a variation, a professional application, of the scientific method of inquiry. That is, it is structured around a hypothesis raising/data gathering procedure. Ideally, insightful questions would determine the appropriate assessment procedures to be used in diagnosis. The better those questions, the more useful the diagnosis.

The validity of reading diagnosis, however, must be determined by longer-range pragmatic evaluation: the degree to which the pupil's reading behavior improves as a result of the diagnostically derived prescriptions. Reading diagnosis as seen in its more formal form, the clinical case study, is a specialized and intensified example of the daily reading assessment/teaching operation, especially if the latter includes criterion-referenced testing and related individualized instruction. In another sense, reading diagnosis may be viewed as a microcosm of the total reading evaluation program, since it includes each of the necessary functions of reading evaluation.

Reading tests are not synonymous with reading diagnosis. Reading tests are not even "diagnostic," as popularly assumed. No reading test can be diagnostic in itself because diagnosis is a dynamic logical process most appropriately employed by a professional, either classroom teacher or reading specialist. Reading tests are useful tools of diagnosis to the extent that they provide needed observations about the subject's reading or reading-related behavior. Most so-called diag-

nostic reading tests are batteries consisting of a number of reading achievement subtests. Often these subtests tap into pivotal or diagnostically sensitive reading behaviors. The battery may provide a mechanical way for using these individual subtest scores to construct a pupil test profile. But it remains for the professional to interpret the profile, and more importantly, to convert this interpretation into appropriate professional action.

The point of this analysis of reading evaluation is this: important and useful as reading tests are, they are no more than cogs in the wheel of reading evaluation. Reading tests are but one form of reading assessment. Reading assessment consists of the broader data-gathering operation of the reading evaluation program. And reading evaluation is but one significant phase of the reading instructional program—an important interdependent phase to be sure, but certainly ancillary to the development of program objectives and curriculum and the selection of instructional tactics and materials. Finally, a reading test or a testing technique itself, for all of the savoir faire which may be used in its development, must be evaluated against a pragmatic criterion—its operational contribution to pupil reading growth through improved reading teaching and program operation. The value of a reading test is highly dependent upon the conversion of its results into effective decisions and instructional action.

Some Evaluative Reactions: Criterion Tests, Performance Contracts, and the Disadvantaged

Reading tests and reading testing perform useful functions within the broader operation of reading evaluation, but they have little educational significance in themselves. This applies also to the specific mechanics and operational problems of reading tests and testing. If they are useful functions or problems of significance, it is because they ultimately bear upon the function of reading evaluation or because they are problematical for reading evaluation. The relative value of criterion and normative referencing, the dilemmas of reading testing the disadvantaged pupil, and the theoretical and practical problems of performance contracting inevitably must be considered in terms of their impact upon reading evaluation.

Criterion-referenced tests. Criterion-referenced testing is hardly a new concept of reading assessment. The direct assessment of a pupil's ability to meet preestablished standards of performance dates back to colonial days, e.g., the saying of the letters, pronouncing the syllabarium, and the oral reading (recoding) of the Lord's Prayer; in short, the catechising of the Horn Book (Smith, 1965). It matters not whether these are acceptable measures of reading according to modern

theory. They were acceptable objectives of colonial day instruction, and they were tested directly by an absolute performance criterion. Illustrative of criterion-referenced reading measures published prior to the common use of norming standards are oral tests, such as the Gray Oral Reading Paragraphs, and certain silent reading tests which assessed an examinee's written summary of the test selections for recall of significant concepts.

It is reasonable to assume that many teachers utilized some form of criterion-referenced assessment prior to the shift in emphasis from oral to silent reading instruction. Many still do, although teachers seem to find that the criterion is less easily referenced for silent reading tasks. Certainly it is less direct than oral. Use of criterion-related reading testing continued in daily instruction even though normative-referenced testing rose to a dominant position in educational testing theory. The term *informal* became an accepted connotation for testing or assessment which was nonstandardized or nonnormative. As used by Gray and others, appropriate use of informal testing involved many of the objectives sought by current criterion-referent testing theory. This functional role was stressed by Gray in 1920:

> Informal tests are tests which are organized by the classroom teacher or supervisor for the purpose of securing accurate records concerning the accomplishments of pupils. They supply a teacher with the facts which are necessary in a scientific organization of her work from day to day. (p. 103)

And reiterated by reading authorities like Durrell (1956):

> Informal tests based upon reading materials used in the classroom and observation of faulty habits and weaknesses in regular instruction provide the best basis for planning classroom instruction. (p. 93)

This does not suggest that the recent trend of support for criterion-referenced measurement is misplaced. It would be unfortunate if a schism should occur among experts in educational measurement, among the reading professional ranks, or between the two over the superiority of normative-referenced vs. criterion-referenced testing. It would be useful if normative tests more nearly reflected the specific objectives of reading programs and produced results which helped to describe the pupil's performance as well as to compare it to peer performance. It can be done. It would be good if the makers of criterion-referenced measures, both classroom teacher and specialist, utilized the lessons learned from nearly five decades of normative test making to improve upon the quality of assessment typically produced by "informal" testing. But in any case, there is a definite need for the differential strengths provided by both sources of reference in reading

evaluation, and test makers should employ the best characteristics of each (Chronbach and Meehl, 1955).

One may remain hopeful yet dubious about the amount of impact criterion-referenced measurement will have upon the quality of reading education and its pupil product, unless some notable improvement takes place in the professional prowess of reading teachers. The publication of criterion-referenced tests developed by reading and testing specialists for broad population use, e.g., the National Assessment Project, should focus attention upon and provide good concrete examples of criterion-referenced testing. Unless sizeable federal grants are available to encourage and reward local school districts for using these tests, their wide permeation of reading programs is unlikely. Such a conclusion seems consistent with observations of the prevailing condition of reading evaluation at the local school level and the improbability of any sudden change in its sensitivity or quality.

It is somewhat questionable whether the local use of nationally developed criterion-referenced measures is a healthy professional direction. Local programs will require locally referenced assessment. Reading teachers and local reading program administrators have a long-established habit of trying to substitute quick, ready-made answers to instructional program problems in place of solution through upgrading individual teacher competency. The most viable use of criterion-referenced assessment is in the adjustment of day-to-day instruction to the specific learning needs of individual children. There is considerable evidence to indicate that most classroom teachers are not able to develop and employ such assessment, regardless of whether it is called criterion-referenced or merely informal. The contribution of criterion-referenced testing to the reading evaluation program will require more than ready-made criterion-referenced tests or even short inservice workshops on "How to Make a Criterion-Referenced Test."

Performance contract testing. The fact that nearly three-quarters of all performance contracts in education are made for reading improvement may contain more than a little irony. For nearly forty years, educators and reading specialists have sold the cruciality of quality reading performance to everyone, with the possible exception of the pupils themselves. In so doing, they impressed parents and boards with the deplorable condition of instruction. And if responsible educators and professional reading groups did not themselves encourage the public to believe that serious reading deficiency was readily remedied through simple instructional programs, they did little to qualify or contradict those who did so. But perhaps the greatest irony of all is that the validity of performance contracting in reading was reinforced by the erroneous belief that reading was one "subject" where pupil growth could be readily ascertained through simple test-

ing rather than by a broader, multiple-phased program of reading evaluation. How could it have been otherwise? Most local reading programs had failed to establish respectable programs of reading evaluation. If the giving of a standardized survey test in reading, usually as a part of the general school achievement test, was good enough for preperformance contracting days, how could such programs argue that it wouldn't serve for a performance contract?

The nature and order of concern exhibited by some professionals about performance contracting appear to be misplaced. Our first concern about a curriculum, a teaching method, or an administrative arrangement should be whether it has a beneficial effect upon pupil learning and behavior. Our least concern as professionals should be with work conditions and job insecurity, particularly since experience has not borne out their reality. Reaction to performance contracts has been too emotional, both pro and con. We need to inject substantial dosages of pragmatism and suspended judgment into our thought stream on this matter. Does the contract produce needed behavioral change? In this, we need some relativity: does the contract system exert more or less educational responsibility than the traditional system? A tough-minded "accountability" analysis of our traditional approaches may prove very enlightening!

It is important that we examine carefully the nature of performance contracting programs presently in existence. It is important that those programs be evaluated by the broadest and most valid means of assessment and data analysis procedures available. But it is just as important to exert our best efforts of total evaluation for every program of reading instruction. The fact that performance contracting in reading has generated considerable soul searching about reading measurement suggests that deficiencies in reading evaluation have existed for some time. The argument that the problem is more serious under performance contracting conditions is fatuous. If performance contracting in reading is as questionable an educational practice as presumed by many professional educators, perhaps the most crucial question is how a well-informed, well-trained, highly-motivated, energetic profession permitted it to get established, both at the local school and the national level? The answer, unlike the question, is hardly academic, though it may be just as obvious.

Reading testing the disadvantaged. Perhaps the thing which impresses one most about the papers of Fitzgibbon and Kasdon is not their apparent differences but their underlying similarities. Both are scholarly, even passionate. If one accepts their differing assumptions, their several positions are well taken. Their most significant semblance, however, is not in substance but in affect; they are laboriously concerned. And, one suspects, both are personally and professionally

troubled about this problem. These papers articulately represent the multifaceted dilemma educators, as well as professional test specialists, face as they try to determine how to test the disadvantaged. The following comments are addressed to the evaluative impact of the dilemma rather than to the numerous technical problems involved.

Admittedly, the horns of this dilemma may be more threatening, more conflicting for the reading evaluator than for the reading instructor, if we can arbitrarily separate the two for sake of argument. Reading teachers can ride in and prick away at the neck muscles of the problem until it is bloody, and if they do this with some appearance of style and sincerity, can ride off with the rationalization that they hacked away in good faith. But the reading measurement person must eventually put his faith to the test. With or without finesse, he must face his moment of truth and either dispatch the beast or leave the arena with the knowledge that he has dishonored his trust. The anxiety manifested by the sincere professional in this business of evaluating the potential, performance, and progress of those populations euphemistically labeled "disadvantaged" becomes understandable. He is caught in a classic approach-avoidance conflict. He wants to hold to the universal principles of scientific measurement which he has inherited from a half-century of research and scientific analysis, but he wishes to avoid the conscience-stricken suspicion that to do so he may be socially insensitive, if not a racial bigot. Of course, psychologists do not insist that the bases for such conflict are necessarily real. It is enough if the professional simply believes they are.

One of the most hallowed tenets of educational measurement and research is that the norming sample and/or the population sampled must be definitively described. If this condition is not met, so the professors have admonished their students, one cannot be sure of the validity of results, let alone draw conclusions from them. So what do we mean by the "disadvantaged"? Are they the members of the lower socioeconomic classes which figured in the studies and literature of the 1940s and 1950s? Are they the "culturally deprived" of the 1960s? Are they the "functionally illiterate" of the 1970s? Are they innercity inhabitants? Are they black? Rich black or poor black? Black boys or black girls? Southern blacks or northern blacks? Southern whites or northern whites? This list of synonyms which have been associated with disadvantagement either directly in the literature or by implication in research studies could be extended indefinitely. A departure from such associative labeling or stereotyping would be some improvement. However, we are not likely to begin to resolve this particular problem until we forgo the use of the general euphemism, "disadvantaged," in favor of definitive description of the population.

This poses a tough task for norm-referenced tests. Not only would

the test maker be expected to include a more definitive and proportional norming sample representative of these population subgroups than he has in the past, but his published results may need to be differentiated into special norms for each representative subpopulation—at least, for those test publishers who suggest that their tests may be particularly useful with such populations. This would not seem a major problem for the criterion-referenced test maker, unless it is his intent to reference his tested behaviors to assumed differences in populations rather than to a direct description of reading performance, regardless of who is doing the reading. In either case, the problem, once again, becomes one of reading evaluation, regardless of the technical complexities it raises for test making.

It is a well-established practice in educational measurement to differentiate between the purposes of achievement testing and those of aptitude testing. An achievement test is concerned with the representative performance of the examinee as an estimate of his present educational accomplishment, knowledge, or skill. An aptitude test, on the other hand, is an estimate of the examinee's potential to profit from appropriate learning experiences, should he be exposed to them. It is understandable that the issues of "culture-free" and "culture-fair" testing should transfer from their traditional source of concern, intelligence testing, to efforts to determine the reading potential of those suspected of experiential disadvantagement. It generally is accepted that a culture-free test of reading capacity is a theoretical ideal which cannot be pragmatically effected. Similarly, it is recognized that tests of intelligence and other measures of reading aptitude may not be culturally fair to those whose linguistic and conceptual backgrounds are inconsistent with the language and concepts utilized in the aptitude test. It would be socially and educationally unfortunate if such results were the basis of decisions not to institute needed adjustments in learning programs for either individuals or groups.

This issue of cultural fairness in testing is very messy, both in theory and application. For example, one could discount its significance, even in the measurement of reading *aptitude,* if present aptitude tests came somewhat near to perfect prediction of reading growth or progress. If they did so, the position might be taken that though it is socially unfortunate that some individuals or groups do not have the background necessary to score well on the test of aptitude, it is very likely that this means they do not have the background, at least at the time of testing, to master that reading instruction to which the aptitude test was predictively highly correlated. However, we have no such assurance. To begin with, most reading aptitude measures have been "congruently" related to reading achievement, not "predictively" validated on reading growth. Even

those studies which have examined the relationship between aptitude scores and reading growth largely have been short run investigations. This problem might be theoretically mitigated if we took care not to confound immediate functional potential to learn from a specific program of instruction with long-range capacity to learn at all. But even this would not provide us with a solid basis for making educational decisions or for educational counseling on a long-term basis. Aptitude test makers and users would do well to remember the lessons learned long ago in the assessment of reading readiness: that readiness results from a compound of variables and that the predictive accuracy of the readiness measure depends heavily upon the nature of the reading instructional program encountered. In short, reading aptitude, like readiness, is better determined through multiple evaluation procedures than through specific test administration.

On the other hand, the accusation that reading achievement tests are culturally unfair would seem wide of its mark. One could accuse some of them of being poorly constructed, standardized, or normed. Of course, reading tests discriminate among examinees. They intend to. There is no point in constructing a normative achievement test that does not discriminate between better readers and poorer readers. Discrimination and reasonable difficulty of items are attributes sought by the test maker. Even in criterion-referenced settings, it would seem empty to construct items or tests to make sure that the examinee gets an acceptable score rather than to determine whether he can perform acceptably on the reading task involved. Reading achievement tests, like other reading, primarily consist of reading the thoughts some writer has encoded. Outside of a few primary-level ocular-motor responses, reading consists of decoding and acting upon linguistic representation of these percepts and concepts, and they usually are drawn from a wide variety of cultural sources. There may be some value in artificially restricting the concepts included in reading tests for young pupils or in a test of basic literacy. But outside of some ego salving, there would seem little value in constructing general reading achievement tests with limited cultural input.

Currently, there seems to be some official effort to develop achievement tests especially made for and normed on innercity schools. The obvious political advantage here is that the better innercity reader will look better than when compared to state and national norms. Such tests would be useful in comparing reading testing results from one innercity school to another. But this could be accomplished by publishing multiple norms on a broadly standardized test which would provide the means whereby both schools and individual students would know how they compare with non-innercity schools and students. Frankly, the only people I have met who look forward to

living and dying in the innercity are those with large incomes, living in well-protected luxury apartments and taking vacations in posh southern spas. The others hope to live elsewhere. Unless we can assume that the innercity student will not compete with others for jobs or for further educational opportunities, unless we are willing to accept responsibility for not encouraging him to expand his personal fulfillment through broader intellectual exposure and personal challenge, it would be flagrantly dishonest of us not to make him aware of what he hasn't mastered and yet needs to learn in order to participate and compete successfully.

A related argument for special tests for disadvantaged pupils may be observed in the request for tests written in what is called the black dialect. *Which* black dialect is seldom indicated. *If* one assumes that it is helpful to initiate reading instruction in terms of the oral language patterns which children bring to school, one can see some value for a temporary use of criterion-referenced measures utilizing the same language. But if these are the same language patterns which are assumed to contribute to educational disadvantagement, we have accepted two logically inconsistent propositions. More importantly, we face the possibility that such practice could reinforce the behavior to be changed. The NAACP took this position in 1971, when it roundly criticized a Ford Foundation project which intended to teach black dialect.

We are just beginning to get the results of careful studies designed to test the much-published assumptions of some linguists that oral language—particularly black language and black dialect--is specifically associated with reading and related learning problems. The findings of Bougere (1968), Johnson (1970), and Melmed (1970), for example, do not support contentions that specific patterns in oral language predicate failure or that a certain dialect differentiates black from nonblack reading performance. More pertinently, such studies do not reveal a clear negative relationship between familiarity with special speaking dialects, polyglot or idiomatic, and success in learning to read (Melmed, 1970). Studies do reveal that reading success is significantly related to experiential impoverishment, in terms of school and general cultural requirements. Increasingly, language scholars emphasize the common generative characteristics of English and the equality of dialect (Deese, 1970). Apparently each of us has a basic dialect and, for only a few, that may be academic or book-type standard English. The important consideration may be the reader's language flexibility rather than his dialect.

The major point here is not that the testing of disadvantaged pupils is either unimportant or impossible. The point is that the prime purpose of testing the disadvantaged is the same as it is for all

pupils—to help devise the best educational program possible. The answer lies not in social-economic rationalizations, ego-salving norms, or the construction of unrealistic reading tests. The answer does not lie in test making and test giving per se. Progress does lie in the development of better total programs of reading evaluation, in evaluation programs which are less concerned with the issue of superiority and more dedicated to providing the specific information needed to improve the learning opportunities of these children and adults.

Evaluation Programs, Teachers, and Professional Accountability

This reaction paper reflects concern with current problems in reading testing. However, the paper has stressed the idea that reading testing is properly an ancillary function of reading evaluation, not an independent entity. The crucial criterion by which a reading test should be judged is its efficacy—its power to contribute useful information to the evaluation of individual or group reading behavior or to the evaluation of the programs responsible for this development. The specific technical problems of reading testing become significant educationally only as they inhibit effective reading evaluation. To illustrate the practicality of this contention, the technical difficulties and philosophical dilemmas pertinent to the three major concerns of the preceding papers can be traced to evaluational assumptions.

The contention of this section is that prevailing difficulties in reading evaluation, as well as in reading testing, are caused by more fundamental professional deficiencies. It is assumed that both reading testing and evaluation are deserving of professional attention and concern to the degree that they contribute to pupil reading growth through reading education programs. There can be little doubt that current reading tests and reading testing practices are in need of improvement. Any such improvement should increase the potential of the reading instructional program. But it is doubtful that improvement in reading tests and testing will have much impact upon reading education unless comprehensive professional action is taken to improve broad practices of reading evaluation per se.

By any accepted criteria, reading evaluation practices across the country are in a deplorable state. In the great majority of school systems they are minimal and haphazard. They seldom are part of a planned systematic program of broad and continuous data gathering. They are geared less to the general and specific objectives of the reading instructional program than to whatever scores a particular standardized reading achievement test provides. They do not reflect the cooperative contributions of the teacher, counselor, administrator, and pupil. The results are fed back into the instructional situation in an improbable manner, and if they are fed back, they seldom extend

to evaluation of the instructional program itself. If current practice in reading testing is analogous to the proverbial halt, then it is being led by the proverbial blind.

It is not enough, however, to improve the administrative practices of the reading evaluation program. No doubt, trained continuing leadership at the local level would produce some change for the better. But no substantial improvement is likely to occur until we cog the wheel of reading evaluation with classroom teachers and reading specialists who understand measurement and evaluation, both in theory and in practice. The classroom teacher should be the pivotal member of the evaluation team, both as a primary source of pupil observation and as the direct instructional effecter of evaluative data. Most professional observers recognize that the typical teacher does not so perform.

Behavior, even that of teachers, usually is the product of multiple determinants. Teacher performance in reading evaluation is related to some extent to those broad factors which bear upon other facets of teacher performance—teacher recruitment, school administrative policies, school resources, pupil population, and the teacher's personal characteristics and problems. Care must be taken not to oversimplify the issue. However, teacher ignorance of the principles and practices of evaluation, reading and otherwise, is well documented. And the problem does not end there. Those who have had the experience of teaching courses in reading measurement and diagnosis at the graduate level can attest to the fact that most teachers are substantially confused about the nature of the reading process itself. It is difficult to see how a teacher can understand the applications of testing and evaluation if she is confused about the behavior she is attempting to assess.

The intent here is not to make the classroom teacher or the reading specialist the heavy in this piece. Many teachers recognize their need to understand more about reading evaluation and the reading process. They are critical of the lack of such substantive preparation in their undergraduate programs. They are dubious of the value of graduate courses in educational measurement and evaluation which stress statistical niceties and ignore the applications of test construction, test interpretation, general assessment procedures, and the principles of evaluation. They are just as dubious about courses in reading education which avoid instruction in the nature of the reading process and how to measure it. Wittingly or unwittingly, teachers may be the tools and partial victims of this educational caper rather than calculated villains. They are better characterized as a "Gang Who Couldn't Shoot Straight" than as Bonnies and Clydes.

But something can and should be done. If scapegoats are needed, none are more deserving than teacher education programs and professional reading associations which have blithely ignored their responsibilities in this matter. A most disturbing element in many discussions of reading testing is the profession's hesitancy to recognize teacher inadequacy as a significant problem needing direct action in itself. In short, there is a general professional unwillingness to admit that our teacher representatives are meandering about in their tactical nakedness. The responsibility for this condition falls more heavily upon reading and elementary teacher education than upon educational measurement. It is time, even if belated, for the reading profession to mount a serious attack upon preparation and certification programs for teachers of reading. The issue of accountability in reading applies to more than performance contracting of instruction.

References

Bougere, M. B. "Selected Factors in Oral Language Related to First Grade Reading Instruction," unpublished doctoral dissertation, University of Chicago, 1968.

Chronbach, L. J., and Meehl, P. E. "Construct Validity in Psychological Tests," *Psychological Bulletin*, 52 (1955), 281-302.

Deese, J. *Psycholinguistics.* Boston: Allyn and Bacon, 1970.

Durrell, D. D. *Improving Reading Instruction.* Yonkers-on-Hudson: World Book, 1956, 93.

Farr, Roger. *Reading: What Can Be Measured?* Newark, Delaware: International Reading Association, 1969.

Gray, William S. "The Value of Informal Tests of Reading Performance," *Journal of Educational Research*, 1 (1920), 103-111.

Johnson, Dora K. "An Investigation of the Oral Language and Oral Reading of Black First Grade Children," unpublished doctoral dissertation, Ohio State University, 1970.

READING TESTS AND THE DISADVANTAGED

Thomas J. Fitzgibbon
Harcourt Brace Jovanovich

When I first started to put this paper together, I did so with trepidation. "What are you doing sticking your neck into that den of test reviewers?" I asked myself. "You know from past experience that in such a situation you can't win. Questioning and probing a test publisher is considerably similar to shooting fish in a barrel." But then I reminded myself that, really, I had accepted the invitation to develop the paper with alacrity, for I saw it as a chance to compare notes and discuss some issues with professional colleagues—persons who are trying to come to grips with many of the same problems that face the test department at Harcourt. What better chance could be found to explore avenues of improvement than with a group such as this?

Thus, I shall deal basically with mutual problems seen through the eyes of a test publisher whose organization has entered into its fiftieth year of endeavoring to research and develop good reading tests for the nation's schools. Some of my views are shared by my colleagues in the test department, some are not. But, no matter whose views, I do hope they will lead to profitable discussion among all.

Now moving directly into the topic of the use of reading tests with the disadvantaged, I should make certain that all know what *disadvantaged* means to me. In educational circles, I believe, the term *disadvantaged* generally refers to those whose environment outside of school does not equip them well to meet the demands of the educational and economic systems of our culture. Stated more explicitly, we can say that those termed disadvantaged are economically poor, are experientially impoverished, live in an environment that is not education-oriented, lack a tradition of literacy, feel rejected by the major cultural groups, suffer from poor self-concepts, and/or have difficulties with the English language considered acceptable by the school.

I should like to add one other emerging characteristic of the disadvantaged—not of the child, but of the community of which he is

a member. That other characteristic is controversy—controversy over the appropriateness of the reading test used, over the fairness or bias, or over whether there should be any testing at all. The community is becoming involved, and in a vital way.

Educational measurement has always been somewhat controversial, not because it was thought unnecessary but because there was lack of agreement as to who should decide upon the questions to be asked. Rarely did this controversy involve the public; rather it was more a disagreement within the educational profession itself. It was the teacher lamenting the time he gave to administering and scoring tests and then "getting nothing he could use" versus the administrator whose motives for testing were on a more global scale and whose reaction to the results often had little impact on the immediate world of the teacher. This situation still exists, but it is now apparent that others are demanding the opportunity to decide which questions should be asked, and they are doing it in a critical vein. For some, tests are, at best, ineffective and, at worst, racist and discriminatory. Pressures—and very effective ones—are being put upon legislators and educators to modify or eliminate present school testing practices. It appears that school people no longer have to worry about involving an apathetic public; now they have to be able to explain and defend what they're doing—sometimes in a highly emotionally charged atmosphere. Accepting this state of affairs, we must ask ourselves: what are some of the things which should be considered when charges of bias and unfairness are made against the reading testing program?

This paper, then, will deal with certain test-related areas which are highly pertinent at this time, and it will do so from a very definite point of view—test sophistication must come to encompass not only a mastery of information about tests and test practices but also a unique blending of good will and political realism. Debate about school tests has become vigorous, even strident, and no longer takes up chief residence in the groves of academe. One doesn't necessarily win the point these days because a test correlates .85 with some measure of achievement; now one must also be prepared to prove that the test is "relevant."

Content and Relevance

It is the content which poses the most problems to mutual understanding and acceptance of tests. In most instances the layman's attack upon a test does not include the term *content validity*. (This is a designator used only by those of us who have had a course in tests and measurements.) More often the attack is leveled at an item or set of items in the tests. To the man in the street (and he doesn't use the term *items* either), the issue is the questions being asked. This would

seem to be an easy criticism to resolve; obviously, the right kind of question is the one which yields a response—correct, incorrect, or representative of one's affective state—which enables the test interpreter to take action. Most of us would agree, I think, that test results are not supposed to "just sit there"; rather they are supposed to help someone to do something because the questions themselves have a redeeming social significance.

To fully comprehend the vehemence with which some tests are rejected, we must also address ourselves to how the question is asked. In most instances, the protestor is objecting more to the way the question is asked than to the objective to which it is linked. This is the area, I believe, which precipitates most charges of bias and unfairness, particularly against standardized achievement tests commonly used in most schools. A good example of this is Wasserman's content critique of the Metropolitan Reading Test used in New York City Schools (1969). Her critique and the rebuttal by Wrightstone (1969) are indicative of the kind of dialogue which can be expected over the issue of how the question is posed to the youngster.

It is very important to understand that Wasserman and Wrightstone are not arguing about *whether* a pupil's word knowledge should be measured but rather *how* it should be questioned. In my opinion, their debate is representative of much of the current controversy about achievement testing. It is not so much a clash of different value systems but, instead, one of mode.

It is time that those of us who are responsible for testing in the schools, whether test publisher or user, pay more attention to what "content validity" means. Cronbach (1969), discusses the logic of evaluation and proposes modifications of earlier thinking as expressed in the 1955 achievement-test version of the *Technical Recommendations for Psychological Tests* (APA, AERA, Second NCME joint committee, 1955). I find his observations to be very helpful to my own thinking and, thus, have included several excerpts from his paper. They are presented out of context and are followed by my comments.

> A content interpretation refers to a universe of tasks or of observations. The universe description is an operational definition that restricts the admissible range of instruments, questions, settings, examiners, etc.; even the narrowest definition defines not a unique operation but a class of operations.
>
> The only indispensable requirement in a universe definition is clarity: Reasonable observers must agree as to what falls within the universe and what is excluded.
>
> In principle, validity of the selection of content is to be judged without considering at all the persons to be tested; attention is restricted to the test materials and the universe description. If the content fits the universe definition, the test is content-valid for persons of all kinds. From

> an absolute point of view the score on a task indicates that the person does or does not possess, in conjunction, *all* the abilities required to perform it successfully. A dictated spelling test is a measure of hearing *and* spelling vocabulary *and* ability to write. In terms of content, however, the spelling test tests ability to spell from dictation. The pupil who is deaf will earn a low score, but that score is a valid report of his inability to spell from dictation. (pp. 23-24)

Comments on the excerpts above. Those who discuss reading tests and their value for youngsters should come to agreement in advance about what they want tested. Thus, if it is agreed that our society puts a premium on knowledge of English word meanings, then the amount of this knowledge present is important to test. If there is no agreement, however, there is no use to argue the merits of a particular test any further. If someone says, or is implying, "I am just as good as you are" and "My culture is as rich in accomplishment and tradition as yours" and "I don't need you" or "I don't want you," there is no way for a test acceptable to this first group to be acceptable to the second. Reading tests are culture specific. On the other hand—even though the two groups may cling to their cultural differences—if there is agreement that certain common skills are necessary for upward mobility in the educational system, and therefore in the power structure, a common reading test will be acceptable. Once this hurdle is passed, we can discuss the set of stimuli (questions), as well as the set of observing operations, and charges against specific items can be handled. When this is done, one usually finds one of two complaints to be paramount: 1) the item will emotionally "turn off" the disadvantaged youngster or 2) the item doesn't relate to the youngster's experiential background. The first complaint, I believe, is increasingly well taken as the level of ethnic pride increases. Most standardized tests are constructed by middle-class people who sometimes clumsily violate the feeling of the test taker without even knowing it. We are giving this complaint a great deal of attention and have moved to meet it by including on our professional staff, both in full time as well as consultant roles, persons who can help us avoid this situation. In a way, I suppose one could say that we have been not so much culture biased as we have been "culture blind."

The second complaint is considerably more difficult to resolve. Here the item writer hasn't insulted anyone; he has asked a question related to a valuable objective but has demanded for response a mastery of symbols a pupil simply doesn't have at his command. "On the vocabulary test, don't ask an innercity kid what a *hostler* is—but if you had just used *hustler,* now that's different." What has happened in this instance is that the protagonists have not really agreed upon the universe; they haven't been specific enough. If they had, the universe

would perhaps have included both words, and each word would have had equal opportunity to appear in the test. This illustration of bias is interestingly similar to one from many communities (including white middle and upper class) whose pupil's average word knowledge scores on a reading test are below community anticipation. Here the feeling is not that the words used are unfair due to home and street environment but, rather, that they are unfair because they were not stressed in the basal reader used by the school system.

In both of these so-called bias situations, one can see the element of threat. The public either did not know what use was to be made of the results; or it did know, but no one asked its opinion; or its opinion was asked but overridden. In any event, there is lack of agreement as to what the testing is all about. This naturally produces anxiety which leads to lack of trust on the part of both groups. The parent attacks the test and the school because his youngster scored poorly, while the educator tries to protect the system which chose the test. If both parent and school had agreed in the first place about fundamental curricular objectives, test results would be better understood and accepted with less anxiety and more equanimity. Then, when results are poor, it would be easier for both parties to admit that sometimes things are pretty bad and some educational changes are needed. Mutual misery, when its dimensions are agreed upon in advance, may serve as the springboard to school and pupil improvement. Now another excerpt from Cronbach:

> Content validity is impermanent. The items or tasks in the test reflect social events, job descriptions, accepted beliefs about the world, decisions about what the curriculum should cover, etc. These change with the passage of time, so that sooner or later the test becomes unrepresentative.
>
> The recommendation that the evaluation battery be comprehensive seems to run counter to the concept that an educational test should measure what has been taught. And students think a test "unfair" when it asks about topics not covered in the course. One can agree that it is unjust to let the fate of an individual be determined by a test that, through no fault of his own, he is ill-prepared for. But this only illustrates once more how a test valid for one decision can be invalid for another.

Comments on the excerpts above. Most administrators I know continually face the problem of including or excluding various youngsters from their system-wide testing programs, and quite often this decision is very difficult to make. A good case in point would be a school system with a Teaching English as a Second Language (TESL) program. Should a first or second grade youngster who must use Spanish as his basic instructional language still be subjected to a first grade reading test in English? In my opinion, English should not

be employed if the major reason for giving the test is for instructional purposes when the instruction is in Spanish. There simply will not be a "match" at this point. On the other hand, the test should perhaps be given if the main purpose is not for teacher use but for an overall system survey, in which test results would be furnished as evidence to a state legislature trying to decide which schools need the most help in improving reading instruction with English as the medium. A test invalid for one decision can be valid for another. I'm quick to point out that I said *perhaps*, mainly because we must first think of the impact upon the youngster. If testing in this instance would simply be another in a long line of failure experiences which have become very uncomfortable to a child, I would get my evidence for the legislature in some other fashion.

The pupil who uses his native language for instructional purposes poses a very unique testing situation. In a curious way, he is like a youngster who has been using the initial teaching alphabet to prepare for later reading without it. His teacher does not wish to delay testing until he is clearly a master of English; she wishes to know how well he is mastering the skills which the medium is helping him to develop. In other words she wishes to know how the process is going, the eventual goal being his mastery of yet another verbal symbol system.

This "in-between" process has received attention at the Southwest Educational Development Laboratory as it has worked with test measures which combine visual and auditory stimuli with a motor or simply graphic response (Randall, 1970). In using these measures, the bilingual status of certain children poses a definite problem in that the receiving language (auditory stimulus) is composed of two overlapping sets. One is in English, and the other is in a different language. In this instance, it appears that stimuli received in either language alone is not sufficient to allow the youngster to finally show all he knows.

Realizing this, the Laboratory started to experiment with use of both languages for stimuli prior to requiring a response, with the resultant finding that the children then scored higher. The sequence in which the languages were presented also was related to scores, depending on the dominance of one language over the other. Higher scores were obtained when the dominant language was presented first. Thus, we have yet another example in the measurement area which shows persons using tests must very clearly know what they are about. And again we see that whether a certain test is content valid may change dramatically over a relatively short period of time.

In sum, content validity is the one characteristic of a test most often challenged by members of minority groups. This may be true because the items, the questions, are so highly visible. In other words, face validity is very important. We have seen, too, that there are

content problems brought about by bilingualism. But there are other reasons for test disapproval, not to say disavowal. Let us examine some of these.

Constructs, Criteria, and Relevance

Charges of unfairness and bias in school tests are also made because 1) there are doubts that what is said to be measured is really being measured and 2) there are fears that the way that tests are used to predict what a pupil or student may do in the future in first grade reading, Algebra I, or freshman year in college is generally detrimental.

To ask what "really is being measured" is to ask about construct validity. When this issue is raised by a minority racial group, it usually means that the group does not believe that the test in question is measuring what the administrator says it is.

"What does it really measure?" is a question which takes on added significance when such ethnic groups as the Puerto Rican, Mexican-American, American-Indian, and Afro-American are demanding their "piece of the action." All of these groups are convinced that education is a tremendously important factor in accelerating upward mobility. What else can their reaction be but anxiety when *Pygmalion in the Classroom* (1968) appears and claims that teacher expectation of pupil behavior could come to serve as a self-fulfilling prophecy—and they know, or fear, that teacher expectations can be shaped by knowledge of test scores?

The charge also is made that these scores, even if they do reflect some kind of ability, are highly susceptible to influences other than the mere presence of the postulated ability (construct). This observation is most certainly true. In interpreting a test score, one must not delude oneself into thinking that its content reflects some absolute or pure trait irrespective of the conditions of measurement or of the population being studied. Unfortunately, simplistic interpretations are frequently made without taking into consideration the multitude of factors which can affect the final outcome. *Guidelines for Testing Minority Group Children* (1964), for example, describes a situation in which a fifth grade achievement test may measure arithmetical knowledge in a middle-class neighborhood where most children are reading up to grade level. However, the same test, with the same content, may be strongly affected by a reading comprehension factor in a school in a disadvantaged area. Thus, the test may be measuring something quite different from what appears to be the case.

Another example of this is seen in the Mexican-American confrontation with the educable mentally retarded (EMR) program in California (1969). In an investigation of placement procedures, Chandler

and Plakas tried to determine whether certain Mexican-American pupils placed in EMR classes really belonged there or whether a language barrier prevented them from being properly assessed as to their ability to perform cognitive tasks. Upon retesting with the Spanish version of the WISC, and using norms developed in Puerto Rico, mean IQ gain was 12.4; 44 of the 47 pupils tested scored higher on the Spanish version. Even more interesting to note is that, after being tested with the Spanish version, only 9 of the 47 scored below the cutoff IQ for placement in EMR classes. I am not certain what intelligence test was used in the first testing and, thus, have no way to speculate validly about the complex of reasons leading to the gain. For my purposes at this point, however, it makes little difference. What I wish to stress is that a question was raised about a "mechanical interpretation" of IQs when the youngster tested came from a non-typical background. In this instance, a language barrier was suggested as the reason for the lower initial scores; in other instances (and, perhaps, even in this one, I do not know), it might as well have been lack of motivation and confidence as the children approached the test-taking situation, or slowness in getting acquainted with what they were being asked to do, or anxiety leading to blocking, or many other factors which could preclude "showing their best."

No matter what a person's best is, two minority-group attitudes have emerged clearly and forcefully. First, whether it is the Mexican-American situation in California or the Hobson versus Hansen case in Washington, D.C. (1967), test results often are seen by the disadvantaged as being used to segregate. Whether the use of a test results in placing a youngster in a certain track or in an EMR class, the charge will quite likely be racism. The second attitude is a corollary to the first; i.e., where the child is placed is seen as a denial of equal educational opportunity and, perhaps, an attempt to keep him down.

Unfortunately, but inevitably, these suspected uses have resulted in wide distrust of testing by many members of minority groups. This is unfortunate because it ultimately hinders research and development which could result in more reliable and valid tests and testing practices. Further, this distrust also has, in some cases, a self-defeating aspect about it. A recent issue of the *New York Times* carried a report that a certain approach to teaching of reading had apparently greatly increased the comprehension ability of a group of ten year olds. I say *apparently* because, being in an experimental situation, those evaluating the new program wanted to look outside its own built-in evaluation process so they could compare the effectiveness of several different approaches to teaching reading skills. But, alas, they could not, since the subdistrict superintendent had previously banned the citywide administered standardized reading test, declaring that such

tests were "useless and had no significance to the learning process." Thus, since they had lost the common denominator of city comparison, they were hampered in weighing the merits of their new reading approach against other new instructional programs being tried out in other places in the city.

Latest witness to this phenomenon is yet another newspaper heading: *Reading Test Hoax Gives School High Official Rank.* The story underneath tells of a deliberate attempt by the school's reading specialist to teach answers to the test items because the test was considered to be unfair. The intent was to use the inflated scores later to expose and dramatize the "evils" of the standardized test. Ironically, the phony scores changed the relative rank upward in such dramatic fashion that, unwittingly, the school was removed from a list of those which had been selected to receive special funds because previously their pupil performance had indicated that extra help was needed.

Attitudes such as those just cited are indicative of the considerable criticism by minority groups of the validity of all educational tests. Critics go so far as to say that since some tests have been misused, all tests must go. This solution is one of clear overkill because it is too generalized and, more importantly, will further deprive the very persons who could benefit most. Let us not, however, miss the most important point of all; i.e., these people are trying to tell us that tests are powerful political and social instruments, and too many times these instruments have been used against them. They are questioning the good will of the test user. It is not my intent at this time to pursue the fostering and nurturing of good will, but neither is it my intent to flee from the issue. As a test maker, I can do something about the matter by becoming more aware of how my work is being used and by furnishing more facts and interpretive aids for its better use. It is true that final responsibility for valid use of test results rests with the person who interprets the results. It is equally true, however, that it is the publisher's responsibility to do his best to prevent damaging test use whether from ignorance or lack of good will.

Much of the criticism of school tests, however, cannot be treated as resulting from charges of unethical use. Rather, I think, it is simply that in our society no one wishes to score at the bottom, or be the last—save, perhaps, the more masochistic of us. In criticism of this nature we find tests being slain as was the courier who brought bad news to the king. It may be that the questions were the right ones for the criterion to be predicted, it's just that the outcomes were too painful; or it may be that the criterion should be examined more closely with an eye to modifying it in some fashion.

This business of the criterion usually makes for rather grim reading, and I'm happy to say that I've found two people who have added a little color to the educational decision-making scene, at least by the titles of their works. Indeed, we owe a debt of gratitude to Astin (1969) for his "Folklore of Selectivity" and to Ristow (1969) for his "Of EMRs and Pterodactyls."

Ristow and Astin both wrote about problems of selection and placement, though they each reported on quite different kinds of populations. Ristow discussed these problems as they affect the California Mexican-American community mentioned earlier, while Astin discussed them in relation to the college-bound student. Essentially, the controversy Ristow reports on arose when the charge was made that a disproportionately large number of minority-group pupils (particularly Mexican-Americans) were being placed in special classes for the educable mentally retarded. (Disproportionate here meant that the number of placements far exceeded what would have been expected from a normal distribution of intelligence.) According to Ristow, there is little doubt that there had been a far greater number of minority-group pupils classed as EMRs than would be expected in the typical situation, but the reason was not misdiagnosis or incorrect placement. In most cases, he suggests that these pupils were correctly diagnosed and placed. His position, however, is more subtle than a simple attempt to maintain the status quo; it has to do with problems of criteria. The major problem is, he says, "... the criteria for the diagnosis and placement of pupils in EMR classes do not distinguish between mental retardation and the lack of qualifications for successful school achievement." He is contending that unless differentiation between these two conditions is made, some youngsters will continue to be misplaced. In making this differentiation, he believes it is a mistake simply to alter the test so that scores on the average will be higher. To quote:

> The purpose of the test is to predict, as accurately as possible, whether or not the pupil will achieve successfully in the school as it presently exists. Obviously, if we alter the school and modify the criteria for successful school achievement, then the test will no longer predict school achievement. If, on the other hand, school achievement remains constant and if we do wish to predict success, then we must not compensate in test administration for any permanent pupil variable which will affect school achievement. The failure of the *Davis-Eells* games is a case in point where the total test was modified to permit high levels of performance but as a result could not predict school success. (p. 6)

This is a very sophisticated argument, as I see it, which finally can be reduced to something like this: modifying a test so that youngsters from minority groups will achieve higher scores will not necessarily

guarantee that they will learn more in the educational program in which they are enrolled; rather, the program itself is what must be changed. Ristow is also pointing out that the test in question will not, by itself, differentiate between those youngsters whose environment has prevented (or, at least, not enhanced) the development of their cognitive abilities and those youngsters who have not had environmental handicaps but who are still very slow in responding to the instructional program.

By no means have I exhausted Ristow's analysis of the EMR situation; however, for my purposes, I now find it expedient to move from "Of EMRs and Pterodactyls" to the "Folklore of Selectivity." Astin, in discussing the challenge of open admissions to higher education, also spends time on tests and criteria. He points out that tests and grade point average (GPA) have been good enough predictors to cause American colleges to use them over and over again. So much so, in fact, that he feels the typical admissions officer today functions as a handicapper at the track; that is, he tries merely to pick winners. Handicappers, Astin stresses, are interested only in predicting the horse's performance—not in improving his performance by trying to make him run better and faster. The problem with this approach is that an educational institution is supposed to function less like a handicapper and more like a jockey or trainer. In other words, the educational institution has a responsibility to improve the performance of the individual, not simply to select those individuals with the greatest potential. In addition, in order to improve performance there will have to be some change in the educational program. This, in turn, will cause other criteria of success to be established which will affect the testing program.

But, let's move quickly to another testing situation which seems light years away from college yet, indeed, proves to have some very similar characteristics—a study reported by Personke and Davis in their "Predictive Validity of English and Spanish Versions of a Readiness Test" (1969). When Spanish-speaking first graders were tested in both English and Spanish with the Metropolitan Readiness Test (MRT) (i.e., the same pupils took the test both in Spanish and English), differences between pupil scores on the test taken in Spanish and in English, in most instances, were found to be slight. Thus, it was suggested that English administration of the MRT probably did not result in inadequate assessment of or testing bias against Spanish-speaking children—at least as far as the language parameter was concerned. This conclusion then led to the question of comparative predictive validities which they felt should be pursued. I will not elaborate on the details of this second phase of the study except to comment that the MRT, Form A, was given twice early in the school year to a group of

youngsters (first grade); one administration in English and the other in Spanish. In May of the same school year, these same 38 children were given, in English only, the Metropolitan Achievement Tests, Primary I Battery. Resulting correlations revealed "striking similarities" between both English and Spanish administrations when the criterion was the reading subtest from the achievement battery. The authors consequently felt this indicated that the readiness test was helpful in predicting certain reading-related achievements for Spanish-speaking first grade pupils; i.e., administration of the MRT in standard English and colloquial Spanish seemed to yield similar predictions, which, in turn, suggested that giving the readiness test in English apparently did not result in test bias.

I have taken pains to report the Personke-Davis studies in some detail because we need to give more thought to the observation they make. One of particular significance which already has been mentioned, I will now reword as follows: test bias, if defined as lowered predictive validity, did not exist in this instance. This is an important observation to make these days because it bears upon the frequently made lay charge that all tests currently used in schools are biased against minority group children (Williams, 1970). It is also important for the professional educator who feels that a test reflecting middle-class values can never be valid when used with youngsters of a lower socioeconomic status. I suspect this is simply not true for all tests for all situations. Just saying this, though, hardly makes the present testing situation more tenable; indeed it will not be more tenable until we start carefully to define the social significance of our criteria; i.e., what we are trying to predict. This is what Personke and Davis are saying, I believe, when they write:

> But it should be noted that most of the children in this study did not learn to read to a measurable extent. The readiness tests, which indicated that the children would not learn to read, must be considered good predictors of reading success as we presently teach reading. It must be asked whether alternative reading programs might have been more effective. This is to say that perhaps the readiness test was valid, but the reading program was not. (p. 84)

And finally:

> It must be understood that this was a small study in number of subjects as well as scope. The evidence demonstrates that the Metropolitan Readiness Tests had a high degree of predictive validity for these subjects. To stop here would be folly. Further studies, with more subjects, should investigate thoroughly the use of this and other tests with culturally diverse children. Program experimentation should be emphasized, and the results carefully measured. It is not enough to note that a test is a valid predictor of success in reading if the prediction for a large group of children is failure. (p. 84)

All of the above has related to predictions and expectations based on those predictions. This leads me to say that I have not made up my mind about the final impact of the concept of self-fulfilling prophecy, except to say that my visceral reaction prompts me to hope that it doesn't cause us to downgrade expectations. I want children to have things expected of them. I am convinced, though, that those harboring the expectations need better information than they have available so that any expectations are as realistic and sensible as possible. For example, the educational situations just presented show the criterion as being more suspect than the test in the decision-making scheme. I deliberately chose to highlight this phenomenon, not to take the onus off testing but rather to express my belief that the behavior being predicted needs questioning and that test and criterion are ultimately intertwined. It is a ham and eggs, Mutt and Jeff, and sometimes even more subtly, an Alphonse and Gaston performance.

Some Observations and Suggestions

The solutions to problems encountered by educational testing as it plays its role in the education of all children of all people are extraordinarily complex. Of the many reasons, there are two I wish to highlight, with the first having to do with the changing relationships within our societal structure. We are seeing and feeling an increased presence, a drive for equal opportunity (with its corollary of "a piece of the action"), and we are witnessing a reaction to this increased presence. Inevitably, the schools become involved because they are looked upon as keys to opportunity and when schools are involved, so is testing. Further, to be involved in these changing relationships means that testing and its impact (whether real or imagined) are often discussed during times of emotional stress—a practice usually not given to the production of reasoned outcomes.

This, then, leads naturally to the second reason why the solutions are complex: those of us who build and research reading tests, and those of us who use the tests in school settings, need to know more about what we're doing and why we're doing it, for there has never been a higher level of interest in educational testing on the part of the public. This interest evidences itself in many diverse ways, ranging from parental queries to classroom teachers, through forced release of citywide achievement test results, to charges that test results are being used to perpetuate segregation. Statements about testing are being made in the most global manner: "All tests must be banned," "There must be a moratorium on testing," or "You can't use those tests validly with kids from this background." These statements are quite often answered or countered by no less sweeping observations: "Listen, when you finally get down to it, you have to be able to

read," "It's not the tests—the kids just don't try," or "So what good did the Davis-Eells do; they didn't do any better on that."

I don't want to belabor this point of increased public interest, but it has much to do with the statement about the solutions to our educational testing problems being complex. For, at the very time we need all the testing competence we can muster in order to discuss intelligently what we are trying to do, we are increasingly being made even more aware of something which has been lurking in the dim recesses of our minds—we have not done an adequate job in training people to understand and use the tests. We are seeing that it is not only newspaper reporters who confuse grade equivalents with percentiles but school administrators as well, and it is not only parents who think an IQ is unchangeable but teachers as well.

In a perverse sort of way, the present challenges to American education are good for testing because one way or another they so often involve test results. (I feel compelled to emphasize *one way or another* since rarely have we seen educational testing in such an ambiguous situation—damned when it does, and damned when it doesn't, depending upon which educational method or theory is being espoused.) The resultant stage front position is causing all of us to examine what we are trying to evaluate as well as why a certain test may or may not be the best way to do the measurement job. There is yet another good coming out of all this, for in trying to understand how test results are affected by the environmental background of a minority-group youngster, we are prompted to increase our concern about environmental impact on *all* youngsters who take all tests. I would venture to say that by being especially sensitive to the problems of testing disadvantaged youngsters, we shall become more thoughtful about any kind of testing we do.

All of this could be interpreted to indicate that the current focus on testing of children from disadvantaged educational backgrounds has encountered new problems, but I must say this is not the case. It has simply brought to light some old slumbering giants that have been with us for a long time. Perhaps those people who kept telling us that "schools give lots of tests, but no one pays much attention to them" were right. Perhaps it didn't make much difference in the past if a teacher, or principal, didn't know how to answer a charge that the tests they gave were "unfair" or "biased" or "racist" and "really didn't measure anything important anyway." But no matter how it might have been, that day is certainly gone and school personnel now must become more competent in explaining what they are trying to accomplish when they test. Test scores are having an increasingly important impact upon our affairs. They are being used by school boards to assign pupils to buildings, by school superintendents to

defend expensive curricular positions, by newspapers as front-page news, and by the President of the United States as he talks of national educational goals. Test results are the currency of the concept of accountability—and are likely to remain so. It seems predictable that they will continue to be used, frequently at key decision-making times which are often times of stress. The prospect is unsettling, for I fear we are not as ready for all this as we should be.

It seems to me that colleagues not infrequently point out my tendency to oversimplify and to follow this with a drive toward closure on some problem or set of problems. Thus, they would not be surprised to find I'm going to do it again. But I'm going to do it again because I am restive and somewhat dismayed. Dismayed because in half a century of standardized educational testing we have not been able to enlist the support of, and inform thoroughly enough, the classroom teacher. That goes for many other types of school personnel as well.

I'm dismayed for another reason which, in effect, is the other side of the coin. With such growing community involvement in school operations, there is an ever-increasing number of questions about tests and how they are used. As a matter of fact, a considerable amount of this involvement has come about because someone has said the schools are "not doing a good job." One way the interested parent tries to satisfy his own question about school adequacy is to ask the opinion of the teacher of his children (and the superintendent, and the counselor, and the school nurse). If the person asked is not sure, or misrepresents through lack of knowledge, the problem is not mitigated but reinforced.

I have discussed controversy and challenge and would like to offer suggestions for further action.

The attempt to measure reading abilities of youngsters from disadvantaged areas has caused some metric problems to surface which always have been present but which are now exacerbated by the situation surrounding the testing. These are not so much problems of content, construct, or predictive validity, but rather problems associated with measurement of gain. The pre-post testing design which has become so prevalent demands, among other things, that the reading test used have several equivalent forms, that it have a plethora of items which function realistically over the range of abilities typical of disadvantaged youngsters, and that it have some sort of continuous score scale ranging across levels of the test. Some of these pre-post testing patterns allow for such a short time interval between testings that any estimate of gain becomes, at worst, dangerous and, at best, ludicrous. Many of these same problems, of course, are associated with performance contracts which have been negotiated with reading in-

struction as the treatment variable. Some people are now trying to call attention to these problems. Stake (1971); Lennon (1971); and Wrightstone, Hogan, and Abbott (1971) come to mind. I hope the fact that this monograph includes a presentation concerning reading tests and performance contracting means that the reading profession feels a responsibility to speak out on these issues. However, speaking out in this instance should not be only within the profession. In addition, and perhaps more importantly, agencies who fund evaluations of reading programs, the independent evaluators who set up the evaluation design, and the school administrators whose schools and pupils are subjected to the entire process must become involved.

What I am about to say next is related to problems of reading testing associated with all students, but it is particularly poignant when we talk of the disadvantaged. It is my belief that not enough teachers know how to teach reading. Too few feel competent when they face their first teaching assignment; moreover, they feel even less competent in the understanding and use of the more commonly used classroom tests. Too few teachers, and perhaps reading clinicians too, have adequate training in educational testing. Too few are competent to choose, employ, and interpret test results to the ultimate betterment of the student's condition. This last statement leads me to make another—there are only a few circumstances under which a student should not be entitled to know about tests he has taken and their results. For children in our disadvantaged category, frequent use of tests as indicators of progress not only can be motivating but also can serve to ease anxieties over test-taking. A combination of two factors has led us to withhold test information: the first is the belief that keeping certain aspects of one's professional operation "sacred" enhances a feeling of trust in one's clients; the second, I believe, is an unwillingness to admit that one is unable to interpret test results with clarity and in such a way that they have meaning to the student, teacher, or community.

The measurement of reading skills of the disadvantaged needs more attention than it is getting; not to the extent that a different set of psychological reading constructs are needed but, rather, in the direction of paying more attention to the inhibitors which impede response to test stimuli. I have in mind here what might be called "test-taking mechanics." It is realistic to assume that most reading testing is of the group, paper-and-pencil type; indeed these have been, in the main, the types of tests that I have been talking about. If this is the case, then we must pay more heed to those who say that many test time limits are too restrictive, that the test is functioning more as one of speed rather than power. Perhaps we shall also find that better orientation to the answer sheet and booklet are needed prior to testing

as well as a larger block of time devoted to solving practice or sample exercises. These seem such small issues, but are they really? In a norm-referenced test, at some points in the scale, it does not take many raw scores to make what may seem a significant change in a derived score.

"Empty calories" has had its' day; now it is time to talk of "empty stanines." The issue here is the providing of reliable differentiation in the usual range of disadvantaged reader scores. With dismaying frequency we find these youngsters piling up at the bottom of the derived score scale, being assigned a stanine 1 label because there's no other place to go, and having their performance interpreted as that which was unnecessary to test because everyone knew they were terrible readers to begin with. This problem can be mitigated by longer reading tests beginning with a fairly large group of items which hover around the .8-.9 difficulty level. It can also be alleviated by preparing more levels within a test battery. This, however, will irritate the constant problem of deciding which level of the test to administer to whom. Again, the trick is to ask the right question to the right pupil at the right time. My only answer to this is that we must redouble our efforts to find ways in which we can help the test administrator decide which is the proper "entry" level into the reading test series. This entire issue is crucially important to acceptance of the testing situation by the youngster, the teacher, and the community. Nothing turns off any of this group more than constant or near failure; nothing raises the cry of irrelevance as surely as this.

We have to devise more reading tests which help the teacher decide what to do next. The survey test, even with the improvements mentioned above, simply cannot ask the number of questions necessary to meet diagnostic criteria in the testing time typically allotted to it. We are finding our diagnostic tests coming into increased demand. This demand, however, places an added responsibility on the consulting staff of the publisher as it works with the teacher toward making testing time more profitable for him. This takes a staff which not only knows what the test can do (and cannot do); it takes one that can help with classroom management and reading materials allocation. Needless to say, staffs like that just don't happen—they have to be built. We are attacking this problem through a series of inhouse intensive workshops in the area of diagnostic reading testing, as well as consummating a number of agreements with university or school reading specialists. We are asking the latter people to help us serve our test customers better and to critique our tests as they see them being used so that the next round of research and development can take advantage of these observations.

All of the above does not negate a real need for good survey reading tests. Administrators and teachers will still need a summative-type test for school status and program evaluation. However, it seems to me that we must explore thoroughly the possibilities of matrix sampling theory for large-scale testing. In some instances, I believe it will allow us to cut testing time and test costs, while still yielding administrative test data of high quality. It might, for example, aid a harassed central office test coordinator in his sometimes thankless task of getting schools in disadvantaged areas to test at all. This approach, of course, does not yield a score for every pupil, but that could be accomplished by increased diagnostic testing.

And what about needs assessment, performance objectives, item banks, and criterion-referenced tests? Are all these things going to help a child we label "disadvantaged"? I think it is possible if lots of time is available and if a structure is devised which will provide stability and continuity of action. A one-year Title III project in a disadvantaged community is not going to do it, but three to five years and community participation might. Needs assessments and resultant direction for reading goals and special tests must be treated in the same vein as the old joke which goes: How do two one-ton porcupines make love? And the answer is: Very carefully. I do not derrogate these efforts—indeed we are involved in them and will continue to be so—but I do urge thoughtful and realistic appraisal of time and money commitments necessary to bring them to fruition.

Joint efforts on the part of test publishers and tax-funded organizations are clearly necessary. It is high time that University X says to Harcourt Brace Jovanovich or vice versa, "this problem is bigger than all of us—let's see what some mutual effort can do." The same should apply to individual school districts, state departments of education, and federal organizations. We do have some modest efforts of this nature underway and are risking some of our time and capital in ventures which we hope will help a youngster we now call disadvantaged. I am quite convinced it will take new alliances to solve some of these difficult reading measurement problems.

And finally, I must stress again—*measurement problems* does not mean only those of a technical nature. The community in which the testing takes place must be listened to, and it must be better informed about why and how its youngsters are being tested. If it is not, disadvantaged children, particularly, may be barred from the aid that good reading testing can give to teacher and school as both go about their very difficult task.

References

APA, AERA, and NCME joint committee. *Standards for Educational and Psychological Tests and Manuals.* Washington, D.C.: American Psychological Association, 1955.

Astin, Alexander W. "Folklore of Selectivity," *Saturday Review*, December 1969, 57-58.

Chandler, John J., and John Plakas. "Spanish Speaking Pupils Classified as Educable Mentally Retarded," *Integrated Education*, 7 (November 1969), 28-33.

Cronbach, Lee J. "Validation of Educational Measures," paper presented at the Educational Testing Service Conference, New York, October 1969, 23-24.

Fishman, Joshua A. et al. "Guidelines for Testing Minority Group Children," *Journal of Social Issues*, 20 (1964).

Horn, T. D. (Ed.). *Reading for the Disadvantaged: Problems of Linguistically Different Learners.* New York: Harcourt, Brace and World, 1970.

Lennon, R. T. "Accountability and Performance Contracting," invited address to the American Educational Research Association, New York City, February 1971.

Lennon, R. T. *Testimony of Dr. Roger T. Lennon as Expert Witness on Psychological Testing.* New York: Harcourt, Brace and World, 1967.

Personke, C. R., and O. L. Davis. "Predictive Validity of English and Spanish Versions of a Readiness Test," *Elementary School Journal*, November 1969, 70-85.

Randall, R. S. "Language Barriers to Measurement," paper read at the Harcourt, Brace and World Invitational Conference on Measurement in Education, Chicago, 1970.

Ristow, L. W. "Of EMRs and Pterodactyls," *Research and Pupil Services Newsletter*, October 1969. Published by the office of the Los Angeles County Superintendent of Schools.

Rosenthal, Robert, and Lenore Jacobson. *Pygmalion in the Classroom.* New York: Holt, Rinehart and Winston, 1968.

Stake, Robert E. "Measuring What Learners Learn," paper prepared with financial support from the National Educational Finance Project and the Office of the Superintendent of Public Instruction, State of Illinois, 1971.

Wasserman, Miriam. "Testing Reading in New York City: A Critique," *Urban Review*, January 1969, 30-35.

Williams, Nevella. "Community Concern about Testing: A Parent's Point of View," in Thomas J. Fitzgibbon (Ed.), *Evaluation in the Inner City.* New York: Harcourt, Brace and World, 1970, 23-37.

Wrightstone, J. Wayne. "As Dr. Wrightstone Sees It," *Urban Review*, September 1969, 45-47.

Wrightstone, J. Wayne, Thomas T. Hogan, and Muriel M. Abbott. "Accountability and Associated Measurement Problems," Test Service Notebook issued by Test Department, Harcourt Brace Jovanovich, 1971.

WHAT IS CRITERION-REFERENCED MEASUREMENT?

Frank B. Womer
University of Michigan

Consider the following questions. One set was designed to be criterion-referenced, one norm-referenced.

1. "My idea went over like a lead balloon with the committee."

   A. From this sentence we can tell that the person's idea was

   - ☐ rejected by the committee.
   - ☐ popular with the committee.
   - ☐ unnoticed by the committee.
   - ☐ I don't know.

   B. The person who said this is probably

   - ☐ bragging.
   - ☐ disappointed.
   - ☐ pleased.
   - ☐ I don't know.

2. In Disneyland, California, there is a street called Main Street, U.S.A. Over one shop on Main Street there is a big sign. It tells us that this is a lock shop. Inside the shop there are all kinds of locks, but they are not for sale. Visitors see *great* locks and tiny locks. Some of the locks are new and others are hundreds of years old. This shop is a lock museum.

   A. The locks in the lockshop are

   - ☐ all old.
   - ☐ all small.
   - ☐ never sold.
   - ☐ never cleaned.

   B. In this story, the word *great* means

   - ☐ nice.
   - ☐ famous.
   - ☐ good.
   - ☐ large.

Perhaps the reader can guess which item is which; perhaps not. It certainly is not obvious which is which. Both items deal with reading; both are multiple-choice; both were designed for students at the middle school level.

The first item is from a Demonstration Package of exercises for the National Assessment Project. It is designed for age 13 students. The second item is from a nationally standardized reading test for grades seven, eight, and nine.

If inspection of items does not provide sufficient cues to differentiate between criterion-referenced and norm-referenced items, what does?

Characteristics of Norm-Referenced and Criterion-Referenced Tests

At this point I probably should present concise definitions of what I think criterion-referenced measurement and norm-referenced measurement are. But there are too many aspects of each which are the same to permit easy differentiation in specific definitions. Rather, I will discuss a series of characteristics that shed light on their similarities and differences and then will attempt to reach a definition, or at least a definitive statement, of what criterion-referenced measurement is.

1. A reasonable starting point in our search for a definition is "intent." What is the major goal of criterion-referenced measurement; what is the major goal of norm-referenced measurement? Criterion-referenced measurement is designed primarily to provide information which can be related easily and meaningfully to specific objectives and specific standards of performance that have been determined independent of the measurement process. Norm-referenced measurement is designed primarily to provide comparative information which can be related to standards that are determined as a part of and are dependent upon the measurement process. Standards for criterion-referenced measurement are primarily external, such as objective standards of performance, or mastery levels; standards for norm-referenced measurement are primarily internal, such as a ranking of students taking a test. The word *primarily* must be emphasized as well as the words *external* and *internal.* Developers of criterion-referenced materials cannot always determine external standards so precisely that they can afford to ignore normative results. Developers of norm-referenced materials have certain standards in mind as items are prepared, standards that do relate to what would be judged good and proper apart from the norms.

Another aspect of intent that tends to differentiate between criterion- and norm-referenced materials is a greater emphasis with

criterion-referenced materials on *describing* performance of individuals and groups rather than on *comparing* their performance. The reverse emphasis is true of norm-referenced materials. Standardized, norm-referenced tests are designed primarily to rank order individuals on the achievement or other characteristic being measured. Criterion-referenced tests are designed primarily to describe the level of performance on the behavior or set of behaviors that a test is measuring and to relate that description to judgments of adequacy.

2. A second point of differentiation is that the criterion-referenced approach puts more emphasis on direct measurement than the norm-referenced approach. One important aspect of this point is that criterion-referenced measurement makes greater use of items that require an individual to "produce" and answer rather than to "recognize" the correctness of an answer. This difference is a matter of degree. Existing items that purport to be both criterion-referenced and norm-referenced are primarily the multiple-choice type, the type which require recognition of an answer. But many developers of criterion-referenced materials, particularly National Assessment, are striving to develop more open-ended items that are designed to measure, as directly as possible, production of an answer or direct demonstration of skill or proficiency. There is no theoretical reason why item format need be different for criterion- or norm-referenced materials. But in practice there are apt to be some differences. It has been demonstrated repeatedly that the multiple-choice format is a most efficient item type for norm-referenced tests. One can control item difficulty rather well in multiple-choice items in a condition that is essential. But it probably is their ease of use that contributes most to their popularity; they can be machine scored.

Criterion-referenced tests, if they are to maximize their relationship to external criteria, should utilize whatever format will produce the best "description" possible. If one's criterion is the memorization of the multiplication tables, the best measurement may consist of items which call for production of products rather than recognition. Whenever production is deemed essential for criterion-referenced measurement, open-ended items are called for rather than choice items.

Proponents of efficiency will argue that a test consisting of multiple-choice items will rank order individuals in essentially the same order as a test of open-ended items. But if one's goal is not ranking, but description, and if one wants to focus attention on how a single item or a small subset of items describes attainment of a specific objective, an open-ended format may be better.

The difference here, then, can be summarized as less reliance on choice-type items for criterion-referenced measurement than for norm-referenced measurement.

3. Another very important difference between criterion- and norm-referenced measurement is in the acceptable limits of difficulty for individual items. The best norm-referenced items are those which have p-values (percentage responding correctly) falling fairly close to a p-value of .50; preferably between .40 and .60. Criterion-referenced items can range in difficulty from .00 to 1.00, since there is no reason to artificially restrict the range of difficulty. In practice, one would not be apt to develop items with p-values of exactly .00 or 1.00, but items with p-values as low as .06 or .08 and as high as .94 or .96 have been reported by National Assessment.

Norm-referenced tests are designed to rank order individuals reliably. Rank orderings with large differences between ranks are more reliable than rank orderings with small differences. One produces a test that yields large differences between ranks by using individual items as close as possible to .50 difficulty.

If one were to use all easy items, p-values around .80 or .90, the resulting test would yield scores closely bunched at the high end of the scale and with minimal variation. If one were to use difficult items only, p-values around .10 or .20, the resulting test would yield scores closely bunched at the low end of the scale and with minimal variation.

Since criterion-referenced tests are designed to describe performance accurately in relation to some external goal(s), it is not necessary to restrict the range of their item difficulty.

If one's goal is to determine how many students know how to spell a given list of ten words, it would be desirable if each student could demonstrate accurate spelling of each word (criterion-referenced). On the other hand, if one wanted to grade spelling achievement, and allow the more able spellers to demonstrate their proficiency, it would be better to have a list of ten words, each of which would have an individual p-value of .40 or .50 or .60 (norm-referenced). This would produce norms with a mean or median close to 5 words correct, with some students answering only a few correct and some getting 9 or 10.

But even this is not a pure distinction. Many items developed for criterion-referenced users will fall into the middle ranges, into the same range of difficulty as those which are most suitable for norm-referenced uses.

A more basic difference between criterion- and norm-referenced measurement is in the establishment of test validity. The basic intent (goal) of norm-referenced measurement is to report behavior as accurately as possible in relation to the performances of peers on the same material. In order to do this, it is essential to build an instrument

which discriminates well between high-scoring and low-scoring individuals. This is done by using items which maximize test variance (middle ranges of difficulty) and which, individually, are answered correctly more often by high-scoring examinees than by low-scoring examinees. Item discrimination is essential.

The basic intent (goal) of criterion-referenced measurement is to describe behavior as accurately as possible in relation to independently pre-established standards of performance deemed important to the user (developer) of the test. This is done by developing items which relate directly, on a judgmental basis, to the standards (criteria) for which the test is designed. Content validity is essential. But validity relates also to test use, not just to test intent and test construction. If one wants primarily to describe behavior in order to assess whether it is satisfactory (for a teacher, for a curriculum specialist), then a criterion-referenced test is more apt to do the job—providing that the standards of performance of the test developer correspond closely to the standards of performance of the prospective user. If a teacher carefully reads all items in a social studies test and says that they sample his own goals of instruction, that test, regardless of norms, is appropriate. If that same teacher "accepts" only two-thirds of the items as being appropriate for his class, that subset of items, *without norms*, is more appropriate for criterion-referenced assessment than the total test, with norms. If, on the other hand, that same teacher is primarily interested in discovering how his students compare to the performance of other social studies students nationally, he must select a norm-referenced social studies test. He should examine competing tests and select the one that comes closest to providing a set of items that are "fair" for his students. He probably will not find a single test in which he approves of every item, but other potential users will be in the same situation. On a reasonably appropriate test, he then can obtain student scores which can be compared to scores of other classes nationally. Whether he has a right to expect average or above-average performance from his students will depend upon a variety of things such as student ability and previous achievement, effectiveness of the class learning situation, etc. If one wants to compare student performance with the performance of other students on a test, the test must have been normed.

There are other ways of looking at validity. If a test user wants a test score which "predicts" behavior on some other current external criterion or on some future external criterion, he needs evidence (usually correlational evidence) that a test does relate well to the other and/or future behavior. Terms sometimes used to describe these situations are *concurrent* and *predictive validity*. Norm-referenced, standardized tests generally are correlated with other external criteria in

order to demonstrate their potential utility. Criterion-referenced tests are seldom concerned with predicting other criteria. They are concerned with accurately sampling and illustrating an existing level of performance.

Users of criterion-referenced tests are apt to use the results in a fashion that would minimize their effectiveness as predictive tools in seeking to develop skills or increase knowledge whenever a test score indicates deficiencies. Even if this were not probable, development of criterion-referenced tests with easy and/or difficult items in them leads to a situation of reduced test variance which works against their utility for predicting other criteria.

Developers of criterion-referenced tests must give first priority to content validity, to the relationship of items to specific objectives. Developers of norm-referenced tests must give first priority to item discrimination and appropriate item difficulty that will maximize test variance. In practice, many criterion-referenced items will be appropriate for norm-referenced tests, and many norm-referenced items will have content validity. Again, it is a matter of degree and of projected test use that differentiate between the two types of tests.

The concepts of test reliability, as they have been carefully developed over the years for standardized norm-referenced tests, cannot be transferred intact to criterion-referenced tests. In norm-referenced tests one is concerned with accuracy (reliability) primarily in the sense of seeking evidence that repetition of a similar, parallel form of a test will produce a similar score, or evidence that if the test is given on a different day, that a similar score will be obtained.

In criterion-referenced tests, one is less concerned with accuracy in the sense of seeking evidence that a pupil's score will remain stable. In fact, one wants to change his score in a positive way. One is more concerned with the adequacy of sampling of the items from whatever domain is being measured, which is only one aspect of reliability in the traditional sense. If a criterion-referenced test is designed to sample a limited domain, a single objective, then an estimate of internal consistency would be useful. Parallel form or test-retest estimates, concerned with stability of performance overtime, have little utility for criterion-referenced testing.

One aspect associated with the assessment of test reliability has different implications for the two types of tests we are concerned with—the question of what to do about guessing on multiple-choice items. Traditional correction for guessing is sometimes used, but many measurement specialists prefer asking examinees to respond to every question so that every one is guessing when he doesn't feel that he knows an answer. This results in scores which are higher than if a correction is performed but which correlate so well with corrected

scores that rankings are almost identical. For norm-referenced tests, this minimizes the problem of guessing.

For criterion-referenced tests, however, the problem is more severe. Since one is primarily concerned with accurately describing knowledge and skills, any element of chance response detracts from an accurate description. National Assessment is faced with this problem in its most severe form because it reports results by individual items, not by groups of items. If one seeks to develop and administer and report individual items that are very difficult (as National Assessment does), there are theoretical limits of difficulty that can be achieved with multiple-choice items. One could posit that by using a four-choice item, it will be impossible to develop items with difficulty less than .25. Yet National Assessment has already reported multiple-choice items with p-values of .20 and .15 and .10 and even less. This is due to two circumstances. National Assessment adds an *I-don't-know* alternative to almost all of its multiple-choice items. This is designed to reduce guessing, and research evidence from National Assessment studies indicates that it does. P-values obtained from multiple-choice exercises with *I-don't-know* added tend to be closer to p-values obtained from the same exercises presented in an open-ended format than do p-values from the same multiple-choice exercises without the *I-don't-know* alternative. A second factor contributing to the low p-values obtained is the popularity of specific distractors (misinformation) that draw many responses.

In my opinion, the guessing problem is important enough with criterion-referenced tests that their developers should use an *I-don't-know* alternative or seriously consider extensive use of open-ended items. This also reflects my belief that in most instances criterion-referenced tests should emphasize *production* of accurate responses rather than *recognition* of accuracy.

4. Still another way to look at the similarities and differences between criterion- and norm-referenced testing is to consider what areas of education are served best by each method. In my opinion, the instructional function in education is better served by criterion-referenced tests. Mastery tests are needed to help determine whether students have attained specific competencies that all are expected to attain. Diagnostic tests are needed to help teachers plan appropriate learning experiences that relate to existing levels of skill development. Both mastery and diagnostic tests should be criterion-referenced. Neither type need to be norm-referenced if the individual items can be defended as direct measures of desired outcomes, and if one focuses on the item rather than on the total score. Mastery tests and diagnostic tests can be normed, of course, if one feels the need of comparing students with students rather than students with standards.

Sometimes a teacher does want to compare his students with others. If so, that requires a norm-referenced test which also has been *approved* by the teacher as a reasonable compilation of items that do sample that teacher's objectives and content.

In my opinion, the guidance function is better served by norm-referenced tests. Most counselors use tests to gain information about a counselee's relative position in relation to other counselees on some aptitude or achievement or attitude.

Counselors are more apt to be looking forward toward potential achievement rather than backward accomplishment. For that purpose, they need to think more of generalized achievements as they relate to the generalized achievements of those with whom a counselee will compete. Predictive validity is of central importance. Norm-referenced tests are the major source of information about predicted performance.

Sometimes, of course, a counselor needs to know whether a counselee has achieved minimum standards in some given subject or area. When that information is needed, a criterion-referenced test will be better.

The assessment or evaluation function of tests may be handled either by norm-referenced or by criterion-referenced tests, depending on the type of information that will be most useful for a given assessment or evaluation. If one wants basically to compare the performance of different schools or different school districts or different methods or materials and if one wants comparative, ranked information that produces maximally reliable differences for a given testing period, then a norm-referenced test is called for. A potential limitation of norm-referenced tests for this purpose is that the comparisons that will be made will be with materials at the middle ranges of difficulty. This is like comparing the typical child in one class with the typical child in another class. It does not provide much useful information about the relative performance of low-achieving students in "basic" skills or knowledges that all are supposed to acquire.

If one wants information that will be maximally useful for assessment or evaluation of mastery of basic skills (defined at some appropriate level of proficiency), then criterion-referenced tests are better. In this situation, one is using a judgmental standard of performance, and all or some or none may reach that level. National Assessment's purposes are of this type, to describe performance at all difficulty levels. A potential limitation of criterion-referenced tests for this purpose is that reliable comparisons of pupils are difficult from a test that allows the inclusion of large numbers of very easy or very difficult items.

5. The historical development of criterion-referenced and norm-referenced measurement is somewhat parallel but certainly not identical. Norm-referenced measurement has its roots in the search for the identification of individual differences in achievement and ability. Psychologists rather than educators were more instrumental in the development of norm-referenced measurement and all of its psychometric implications. This is not to discount the contributions of educators, but psychologists were more active in the process. As early as 1903, Binet developed a test designed to separate the mentally retarded from the normal child. From the beginning, and still today, the emphasis was on reliably identifying individual differences.

Criterion-referenced measurement can be traced to the early concern of some educators with the desirability of individualized instruction and of the desirability of relating instruction and evaluation to objectives. The post World War I era saw initiation of these developments. The Winnetka plan of 1925 combined these concerns in the use of clearly specified objectives in a fashion that helped guide individual-learning sequences. This was an example of what we now call educational evaluation.

Through the 1930s, 40s, and 50s, both movements continued, but standardized testing as it is known today was and is primarily a product of the individual differences approach. The work of Tyler in evaluation and of Skinner in programed instruction and of others began to lead some educators to a feeling that norm-referenced standardized testing contributed little to their needs for assessing mastery of academic work as it relates to stated objectives.

In the early 1960s, Glaser began using the terms *norm-referenced* and *criterion-referenced* to differentiate between the activities and procedures that had been developed through the years by those attempting to measure individual differences and those attempting to measure attainment of educational objectives. The two movements have been interrelated for many years; no doubt they will continue to be interrelated.

Definition of Criterion-Referenced Measurement

This rather extensive look at similarities and differences between criterion- and norm-referenced testing has led us to the point of formulating a definition of criterion-referenced measurement. If one accepts the points made in this paper, then a criterion-referenced test is one which is designed to provide information about attainment of a specific objective (criterion), which emphasizes direct measurement through the use of differing formats, which may use items at varying difficulty levels, which must have content validity, which must be

internally consistent (or report items separately), which must minimize guessing, and which is particularly useful for instructional and evaluative purposes.

It should be noted that each statement of differences that I see between criterion- and norm-referenced tests is qualified with a potential exception. This is because I see the two types of tests as positions on a continuum rather than as separate categories. Just as ability and achievement measurements differ at some points, they overlap at others. The best norm-referenced tests will contain items that are closely tied to desired outcomes, objectives, or other criteria. The best criterion-referenced tests will furnish item statistics that help to provide normative information.

In stressing the differences between criterion- and norm-referenced tests, we should not lose sight of the similarities that exist; but in qualifying the differences by citing conflicting or overlapping examples, we must not assume that the differences are semantic only. Criterion-referenced tests *are* different from norm-referenced tests, and they have a place in educational measurement. They should take that place by providing information not readily available from other sources to supplement but not replace existing methodology.

CRITERION-REFERENCED TESTS: A CRITIQUE

Frederick B. Davis
University of Pennsylvania

It has become evident during the past few years that many educators and psychometricians have been confused about the purposes and characteristics of what have been called criterion-referenced tests and about how they differ from norm-referenced tests. Three of the more common misconceptions are:

1. The belief that criterion-referenced tests have been carefully constructed to measure performance in the elements of skill and knowledge that are the objectives of defined instructional units while norm-referenced tests have not been carefully constructed to do the same thing.
2. The belief that criterion-referenced tests must be used with a specified "passing mark" that separates those who are considered to have mastered the content tested from those who are not.
3. The belief that long-accepted principles of test theory do not cover the particular requirements of constructing criterion-referenced tests or of estimating their validity or accuracy of measurement.

Let us consider the purpose and the essential distinguishing characteristics of criterion-referenced tests as these have been set forth by Glaser and Klaus (1962), Glaser (1963), and Glaser and Nitko (1970). The purpose was stated succinctly by Nitko (1970, p. 38): "A criterion-referenced test is one that is deliberately constructed to give scores that tell what kinds of behaviors individuals with those scores can demonstrate." Clearly, this statement implies that criterion-referenced tests are intended to be used as diagnostic instruments for identifying highly specific behaviors that examinees can or cannot perform. It also implies that great care must be exercised in drawing up the outline or plan for a criterion-referenced test in order to make

sure that a representative sample of all of the behaviors that the objectives of instruction call for is measured by the items. In short, the content validity of the test is to be guaranteed by a detailed analysis of the objectives of instruction and by the writing of items that elicit examinee behaviors that literally constitute overt manifestations of the feelings, skills, and knowledge (facts and understandings) that make up the objectives. Glaser and Nitko (1970, p. 653) defined a criterion-referenced test as a

> measuring instrument deliberately constructed to yield measurements that are directly interpretable as performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain and such measurements are directly referenced to this domain for each individual measured.

Nothing in this definition precludes such tests from covering rather wide domains with items that elicit behaviors properly representing all of the objectives in the domain. In fact, Nitko (1970, p. 38) illustrated the interpretation of a score of 30 derived from a criterion-referenced test covering a rather wide domain: namely, elementary school geometry. He wrote:

> . . . a score of 30 might mean that, along with a number of lower behaviors, the student is able to identify pictures of open continuous curves, lines, line segments, and rays; can state how these are related to each other; and can write symbolic names for specific illustrations of them. He can identify pictures of intersecting and nonintersecting lines and can name the point of intersection. This score would also mean that the student could *not* demonstrate high-level behaviors, such as identifying pictures that show angles; naming angles with three points; identifying the vertex of a triangle and an angle; identifying perpendicular lines; using a compass for bisection or drawing perpendiculars; and so on.

This interpretation is based on the assumption that the behaviors have been arranged in a hierarchy of complexity and difficulty so that if a given examinee gets a score of 30, he will have passed each of the first 30 items and will have failed all subsequent items. For this to happen in the case of every examinee who obtains a score of 30, the tetrachoric intercorrelations of the items must be unity, and the rank ordering of difficulty of the items must be the same for every examinee. Needless to say, these conditions are never met in actual practice. To the extent that the tetrachoric intercorrelations of the items are less than perfect, scores of 30 obtained by different examinees will represent the result of marking correctly somewhat different sets of 30 items in the test. Consequently, Nitko's interpretation is not appropriate for the type of test that he describes, which is simply a survey

test measuring achievement by means of items judged subjectively to elicit a representative sample of the diverse behaviors that make up the content to be covered.

Two Legitimate Interpretations

If we limit ourselves to content-referenced interpretations of scores from a carefully constructed survey test, two legitimate types of interpretations can be made:

1. We may estimate the percent of the behaviors in the domain that the examinees have shown that they can perform correctly. If multiple-choice items are used, scores on the test that have been corrected for chance success will ordinarily allow making a better estimate of this percent than will number-right scores. It should be noted, however, that this type of content-referenced interpretation does not indicate the particular behaviors that have or have not been demonstrated by each examinee. Therefore, it does not fulfill the purpose of criterion-referenced tests stated by Nitko (1970, p. 38).
2. We can determine whether any examinee did or did not correctly demonstrate the specific behavior tested by each separate item. But it is dangerous to infer that the examinee's performance would be at the same level of competence on each of a large number of equivalent (though not identical) items testing the same behavior. Although the best estimate of his true level of competence with respect to a specific behavior is his score on the one item testing it that he has tried, this estimate is subject to error, possibly to a far greater degree of error than we ordinarily tolerate in test interpretation. Unless satisfactory evidence to the contrary is provided, diagnosis of individual strengths and weaknesses on the basis of one-item tests should be regarded as highly tentative.

Theoretically, the accuracy of measurement of a one-item test for any given examinee could be estimated by obtaining the standard deviation of scores on a large number of equivalent (though not identical) items administered to him under specified conditions. The standard deviation of these scores would be the standard error of measurement of that individual's obtained scores. In practice, we are unable to administer a sufficiently large number of equivalent items to any one individual, so we may administer two equivalent items to a large number of examinees and compute the overall standard error of measurement as an estimate of the standard error of measurement of the obtained single-item score of any examinee drawn at random from the

sample. The required equation for an item scored *1* for a correct response and *0* for an incorrect response or an omission is:

$$s_{meas\ i} = \sqrt{p_i q_i\ (1 - r_{iI})}$$

where p_i = the proportion of the sample that marked the item correctly; $q_i = 1 - p_i$; and r_{iI} = the product-moment correlation coefficient between scores on the two equivalent items. Clearly, the product $p_i q_i$ is largest for items of 50 percent difficulty in a sample (.50 x .50 = .25) and becomes small for difficult or easy items for example, when $p_i = .90$, $p_i q_i = .09$. Since criterion-referenced tests are often administered immediately after a unit of material has been taught to find out what behaviors have or have not been learned by each pupil, the items of which they are made up are usually found to be easy. Ordinarily, the reliability coefficients of single items are very small, ranging from, say, .10 to .20. The writer found in a sample of 998 high-school seniors that the median reliability coefficients of very homogeneous perceptual items ranged from about .15 to .18. Yet Scandura and Durnin (1971) report data indicating that the reliability coefficients of single items testing highly specific behaviors (pertaining to the use of rules in solving arithmetic problems) that had been taught and practiced just prior to the testing were as high as .60 to .90 in very small samples. It may be that under certain special circumstances, single test items have higher reliability coefficients than would be expected.

Scored *1* for a correct response and *0* for an incorrect response or an omission, a single item that was answered correctly by 68 percent of a sample and that had a reliability coefficient of .75 would have a standard error of measurement of about .23. Therefore, an examinee who obtained a score of *1* would be unlikely by chance alone to obtain a score of *0* on an equivalent item. If this item displayed a reliability coefficient of .15, however, it would have a standard error of measurement of .43. Under these circumstances, an examinee who obtained a score of *1* could fairly readily obtain a score of *0* by chance alone on an equivalent item. Additional experimental evidence is needed to determine the standard errors of measurement of short diagnostic tests administered directly after the content measured by the tests has been taught.

From this discussion it is apparent that, although the second type of content-referenced interpretation does indicate the particular behaviors that any examinee has demonstrated, such data may be so unreliable as to make them of doubtful value. Twenty or thirty years ago some test-scoring services reported results in such a way that pupils and teachers could see exactly which items in achievement tests had been marked correctly or incorrectly. But these data have not

become widely used, partly because they publicized the scoring keys for the tests and partly because they were unreliable.

What Is Needed?

Altogether, we may conclude that a criterion-referenced test that covers a wide domain is not likely to provide data that satisfactorily fulfill the basic purpose of such tests. What is needed is a coordinated set of diagnostic tests, each of which is made up of items that are homogeneous in the sense that they test performance on one specific behavior or on a cluster of behaviors that are taught as a unit. The experimental justification for obtaining a single total score from items measuring a cluster of behaviors would consist of evidence that the tetrachoric intercorrelations of single items in the cluster are as high, or almost as high, as their reliability coefficients would permit and lower than their correlations with single items in other diagnostic tests in the coordinated set covering the domain being measured. Each test in a set would comprise enough items so that a perfect score on it would not likely be obtained by chance alone (at some designated level of probability) by an examinee who had not truly mastered the behavior being tested.

By this time, it is apparent that, from the point of view of classical test theory, criterion-referenced tests are simply achievement tests carefully constructed (as all achievement tests should be) to make their constituent items measure a representative sample of all of the behaviors in the domain to be tested. Their essential distinguishing characteristic lies not in the tests themselves but in the fact that only content-referenced interpretations are to be made of their scores. For this reason, their scores need not be expressed in units that approximate equal intervals; neither must percentile ranks and norms be provided for them although, in defined samples of adequate size, they easily could be.

The definition of criterion-referenced tests given by Glaser and Nitko (1970, p. 653) covers both survey and diagnostic tests, but content-referenced interpretations of survey tests are, as already noted, unlikely to fulfill adequately the main purpose (Nitko, 1970, p. 38) that such tests are intended to serve. Hence, a restriction on the homogeneity of the content should be added to the definition of criterion-referenced tests. The writer suggests the following:

> A criterion-referenced test is a diagnostic achievement test deliberately constructed to yield accurate individual scores that are to be interpreted solely in terms of the content tested. The latter must cover only one specific behavior or a cluster of demonstrably homogeneous behaviors that are taught as a unit.

In conclusion, a brief discussion of the role of item-analysis procedures in the development of criterion-referenced tests is appropriate. Naturally, for informal classroom tests that are intended for use only once or twice in small groups, the cost and effort involved in getting item-analysis data cannot be justified; but for preliminary versions of criterion-referenced tests that are to be used in final form with curriculum materials designed for widespread use, they can be very useful if they are obtained in a large sample representative of the population in which the test is to be used. With high-speed computing equipment, the tetrachoric intercorrelations of the items can readily be computed and, for multiple-choice items, tabulations of the percent of examinees who marked each choice (or no choice) in a high-scoring subsample and a low-scoring subsample can be prepared. The latter are likely to prove useful in detecting items that are clearly defective in some unexpected way. Insightful revision or elimination of items is the most important outcome of item analysis. Unfortunately, item-analysis data have often been used mechanically to select items solely on the basis of item test correlation coefficients of one sort or another. As Davis (1952) pointed out,

> For achievement tests, great care must be exercised that items judged unacceptable by subject-matter experts be excluded and that the final form preserve the balance among topics specified in the test outline. Then, too, proper regard for the shape of the distribution of item difficulties must be observed, as noted earlier in this article. The value of item-discrimination indices must always be considered in the light of the adequacy of the criterion variable, the purpose for which the test is to be used, and the way it serves that purpose . . . . the usefulness of item-discrimination indices is often smaller than is commonly supposed. (pp. 116-118)

Like discrimination indices, difficulty indices have often been misused. For example, items close to 50 percent difficulty have frequently been selected for a test in the belief that such items are perfectly pitched in difficulty. But for tests made up of more than one item, this is true only when it is desired to maximize the number of differentiations that can be made among all of the examinees when the product-moment intercorrelations of the items average .33 or lower (as they ordinarily do) or when it is desired to maximize the number of differentiations that can be made between examinees below and examinees above the raw-score median regardless of the level of item intercorrelation. Since neither of these objectives is likely to be relevant in the development of diagnostic achievement tests designed to provide content-referenced interpretations, classical test theory suggests that items should *not* be selected on this basis. In fact, it suggests that item-difficulty levels need not be used in constructing

criterion-referenced tests; item difficulty should come about simply as a by-product of efforts to make the items elicit behaviors that constitute overt manifestations of the feelings, skills, and knowledge that make up the objectives of instruction and of the effectiveness of the procedures used to teach these objectives.

The writer hopes that this brief paper will lead to a better understanding of the nature of criterion-referenced tests and of their place in the broad spectrum of measuring instruments that have been developed over the past half century. He hopes that it will encourage the development of tests that yield scores for which valid and reliable content-referenced interpretations may be made.

References

Davis, F. B. "Item Analysis in Relation to Educational and Psychological Testing," *Psychological Bulletin*, 49 (1952), 97-121.

Davis, F. B. "Research in Comprehension in Reading," *Reading Research Quarterly*, 3 (1968), 499-545.

Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes," *American Psychologist*, 18 (1963), 519-521.

Glazer, R., and D. J. Klaus. "Proficiency Measurement: Assessing Human Performance," in R. M. Gagné (Ed.), *Psychological Principles in Systems Development*. New York: Holt, Rinehart and Winston, 1962, 419-474.

Glaser, R., and A. J. Nitko. "Measurement in Learning and Instruction," in R. L. Thorndike (Ed.), *Educational Measurement*. Washington: American Council on Education, 1970, 625-670.

Nitko, A. J. "Criterion-Referenced Testing in the Context of Instruction," *Testing in Turmoil: A Conference on Problems and Issues in Educational Measurement*. Greenwich, Connecticut: Educational Records Bureau, 1970, 37-40.

Scandura, J. M., and J. H. Durnin. *Assessing Behavior Potential: Adequacy of Basic Theoretical Assumptions*. Philadelphia: University of Pennsylvania, 1971. (Mimeo)

READING TESTS AND PERFORMANCE CONTRACTING

Thomas P. Hogan
University of Wisconsin at Green Bay

The following questions serve as guideposts for a discussion of the relationship between reading and performance contracting: What are the particular problems of using reading tests in a performance contract? What particular demands does a performance contract put on a reading test? What specific recommendations can be made to the school administrator and performance contractor in using reading tests in a performance contract?

Performance Contracts

Definition of a Performance Contract

Virtually everyone within education has some understanding of what a performance contract is. This understanding is usually verbalized in terms of the current series of contracts funded by the United States Office of Education and Office for Economic Opportunity. The most famous—or should we say infamous—of these was the Texarkana Project, which burst upon the educational scene in 1969-1970. Some twenty contracts, modeled on the Texarkana Project, were in effect during the 1970-1971 school year. The formal structure of these contracts is generally incorporated into the public's conception of what is meant by a performance contract. That formal structure has been outlined, as well as promulgated, most explicitly by Lessinger (1970). The structure includes the following salient features: 1) The contract is written with an educational materials supplier, 2) performance standards are agreed upon and objectively measured, usually by some nationally standardized test, 3) the contracting supplier derives monetary rewards for successful execution and/or penalties for unsuccessful execution of the contract, and 4) the target population is ordinarily a disadvantaged or academically below average group. In addition to these features, the formal apparatus of the performance contract includes a number of less well known characteristics such as the pre-

audit, the management support group, and the RFP (request for proposals).

Very little imagination is required to see that the formal structure of the Lessinger-type contract is only one of several potential types of academic performance contracts. For example, it is certainly not necessary that the contract be written with an educational materials supplier. A contract might be written with a teacher, a principal, a university team, or any other person or group that thinks it can lead the students to the desired goal. Of course, school officials must have reasonable assurance that the contractor has a reasonable chance of succeeding, in order to protect the interests of the children. The contract does not have to specify rewards and penalties. The target population need not be a disadvantaged group. In fact, there has already been at least one instance where this restriction has been removed—somewhat by accident (Mecklenburg and Wilson, 1971). It would certainly be possible to write a contract without using nationally standardized tests, or any tests at all for that matter.

The irreducible elements of the performance contracting concept seem to be the following. First, some party (e.g., corporation, teacher) accepts explicit responsibility for the project. Second, exactly which students are involved is specified (e.g., all students in a given classroom or all students in a state who are below average in reading). Third, goals or outcomes are formulated in terms of students' behavior (what students will be able to do as a result of the educational program, not what will be done to them in the program). For example, students will be able to read at a certain level or they will *not* drop out of school. Fourth, methods for determining whether objectives have been met are clearly agreed upon. Finally, all of the latter conditions are voluntarily agreed to by school officials (or parents?) and the party mentioned in the first condition, with the agreement and conditions probably specified in writing.

With the momentum now gained by the performance contracting phenomenon, it is reasonable to predict that a great variety of proposals meeting the latter five conditions will arise within the next several years. Many of these proposals will only vaguely resemble the Lessinger model.

Within the context of this paper, attention is focused on this more general model incorporating the latter five conditions. Discussion is not limited to the Lessinger model, although it is covered. Questions about the use of reading tests arise in the context of the third and fourth conditions, i.e., the goals to be specified and the evaluation of the attainment of those goals.

Performance contracting has been a topic of heated, often bitter, controversy during its short history. On the one hand, proponents of

contracting, particularly within governmental agencies, have proclaimed the contract as education's solution to the haphazard, often-stumbling process of educating our children. On the other hand, teachers, particularly as represented by teachers' unions, have charged that contracting is a hoax, a profit-mongering scheme in which children are mere pawns.

Since the performance contracting phenomenon is so controversial, it may be helpful to the reader if I lay bare my own biases regarding the matter. Those biases include the following. First, I predict that ten years from now we will look back on performance contracting as an educational fad. Second, the demise of performance contracting will not be caused by any fundamental fault in the concept but by a) the fact that many contracts are based on the use of unsound educational practices and materials and b) the expense and complication entailed in contracting, as presently conceived, become unbearable. Third, I have little sympathy for the charges that contracting is simply profiteering. Education is in fact a business that involves large chunks of cash—and the largest chunks go to precisely those people who are making charges regarding profiteering at the expense of children. The last and most pervasive of my biases is that I acknowledge my predictions as such. They may not be true. Performance contracting should be given a fair chance. It is a new approach that may have merit. The remainder of my remarks are intended to help give it a fair chance, from the point of view of the school administrator, the contractor, and the student.

General Notes on Contracts

Before analyzing the specific problems of reading tests *vis a vis* the general model for performance contracting, it would be helpful to discuss two special topics which are only obliquely related to those specific problems but are quite important for gaining perspective on the entire subject. One of these topics deals with the nature of contracts in general. The other describes a problem peculiar to the initial efforts in establishing performance contracts.

1. One of the most important thrusts of the current contracting phenomenon has nothing to do with the structure of the contract. It is rather an underlying thesis which might be expressed syllogistically as follows. Big business is run on a contracting basis. Education is big business. Therefore, education ought to be run on a contracting basis. Hours could be spent discussing this argument. Here, however, only the following observations are made. When a contract (in big business) is written to cover a very complex phenomenon, it is virtually impossible to spell out all possible contin-

gencies. Except for the simplest types of contracts, certain factors will have to be negotiated after completion of the contract period. Such negotiations depend partially on the good will of the two parties. Further, many contracts are violated without penalty, the most familiar examples being certain labor contracts and defense contracts. The latter involve not only missed deadlines (performance objectives) but actual increases rather than reductions in payment. These remarks are not intended as an indictment of contracts, either for education or business. Contracts are very useful instruments. But they are not the idyllic, fool-proof system portrayed by some proponents of educational contracting.

2. Much attention will be focused on the outcomes of the various contracts now in effect throughout the country in the OEO project. The outcomes of these projects will have a substantial impact on public opinion regarding performance contracting. It appears safe to predict that considerable controversy will surround the test results for these projects. The controversy will center around questionable test selection, administration, and interpretation. Accusations regarding questionable practices will frequently be true. To a large extent, blame for this situation may be laid at Washington's doorstep for the great haste in which initial evaluation efforts for the bevy of current contracts had to be executed. The subcontract for independent educational accomplishment audit (IEAA) of the projects was not signed until approximately one week prior to the mandated pretest date. Allowing one week for securing test materials from warehouses, distributing the materials across the country, training administrators, and administering the test is sheer nonsense. A recent personal experience further illustrates the point. Several weeks ago I received a telephone call late at night from an auditor in a distant state. His contract called for reporting all results in terms of grade equivalents, even at the kindergarten level. The contract had been signed in such haste that no one bothered to investigate whether the test to be used had grade equivalents at the kindergarten level. And, he asked, would I please explain how to derive grade equivalents for kindergarten—and below!

Ironically, Washington's haste in attempting to introduce performance contracting to the nation's schools may be precisely the factor that will destroy the schools' confidence in the concept. (Perhaps OE and OEO need a Management Support Group.) Judgments, either pro or con, regarding performance contracting during its first year or two of widespread use should be very tentative.

Statement of Goals

As indicated previously, the relationship between reading tests and performance contracting becomes critical in the light of the goals specified for students and the methods for determining attainment of those goals. The contracting phenomenon has laid great emphasis on specifying goals clearly and measuring attainment objectively. Too little emphasis has been placed on specifying goals appropriately and measuring their attainment validly.

The selection of a reading test will obviously depend on the nature of the goals specified. It is essential, therefore, that the nature of possible goals be examined. From the point of view of reading tests, goals may be categorized along three dimensions.

First, the contract may specify fixed or variable goals. A fixed goal refers to some absolute level of performance which must be reached by students. For example, the contract may call for bringing all students "to grade level" in reading or "to within one year of grade level." Or students may be required to attain a certain score on a criterion-referenced test. A variable goal is one in which the final level to be attained depends on the initial status of the student or group of students. There are two commonly used types of variable goals. One sets the goal in terms of growth. For example, a contract may call for a year's growth in reading for a year's instruction. A second type of variable goal is based on expectancy data, e.g., expected reading performance for various levels of IQ or for different socioeconomic levels. A more detailed discussion of fixed and variable goals may be found in Hogan (1971).

Second, goals may vary by referring either directly or indirectly to reading skill. A direct goal is one which refers directly to some reading ability such as comprehension. An indirect goal refers to some behavior which is presumed to be related to reading skill but which does not directly involve that skill. For example, a contract might call for increasing the frequency with which students visit the library on a voluntary basis. Going to the library is not a reading skill, but we may presume that it is at least one favorable outcome of a good reading program. The various "dropout" projects are based on indirect goals. The purpose of school is not simply to keep students in school. However, decreasing the number of dropouts may be a beneficial effect of an educational program.

Third, goals may be differentiated in terms of the curricular content involved. This is an exceptionally critical distinction for purposes of the present discussion. No exhaustive list of categories can be given for this dimension. The following categories suggest the kinds of distinctions which are intended: vocabulary or word reading, paragraph

comprehension, decoding skills, rate of reading, and enjoyment of reading.

The local school must carefully consider each of these dimensions and select goals which meet its particular needs.

Time of Test Selection

Once goals have been established, and there is no guarantee that they will be specified clearly, the test for determining attainment of the goals must be selected. Since meeting the basic terms of the contract depends on test results, importance of test selection can hardly be overemphasized. It is essential that a test be selected which will correspond with the enunciated goals of the contract. Too frequently, a contract calls for a specified level of performance or rate of growth simply in terms of "reading." But what is reading? Is it paragraph comprehension, decoding skills, word reading, or what? Often, the answer to this question is not forthcoming until after the contract has been signed. A reading test is selected in advance, only to find that the test is composed of several different parts. Then an ad hoc decision must be made about what parts of the test will be used. This procedure is highly unsatisfactory. Exactly what kinds of tests or parts of a given test are to be used must be agreed upon in advance. Statements regarding this matter should be included in the contract.

Kinds of Tests

One of the most basic questions to be answered is what kind of test should be used. Three kinds of tests may be considered under this question. First, there is the familiar norm-referenced test which provides national norms. Second, there are criterion-referenced tests which provide absolute standards by which performance may be judged. Finally, there are tests specifically designed for certain instructional materials, e.g., end-of-unit tests for reading textbooks. Which of these types of tests should be used?

Most, perhaps all, performance contracts written to date have depended on the norm-referenced type of test. These are the well-known and widely used reading tests such as those found in the five or six major achievement batteries. In light of the oft-voiced criticism of these tests, the widespread acceptance of them in performance contracting seems surprising. On the other hand, this acceptance is not so surprising. These tests provide a wealth of data, not only in terms of national norms but also in terms of reliability, relationships with other tests, performance by ability level, and many other matters. They represent educators' best efforts to produce valid, reliable, convenient

measures of reading performance. Further, they have wide acceptance not only among educators but also among laymen.

Much has been written in the past several years about the merits of criterion-referenced tests. Such tests are said to be particularly appropriate for accountability purposes and, by implication, for performance contracts (Millman, 1970; Tyler, 1970). If the promise of criterion-referenced testing is even partially correct—and I happen to believe it is—why aren't these types of tests used in performance contracting? There are two answers to this question. First, many of the demands of performance contracts are currently phrased in normative terminology; e.g., students will read at or above national norms. The concepts of norm-referenced testing are so well entrenched in the public's consciousness that such terminology will continue to be used for many years to come, regardless of the progress made in the criterion-referenced field.

Second, and more important, well-developed criterion-referenced tests are simply not available today. And, in the reading field, they are not likely to be available any sooner than about 1976, if then. The concepts of criterion referencing are easily applied to areas such as arithmetic computation. But reading is quite a different matter. What is adequate understanding of a paragraph? What kinds of connected, written discourse should be understood? Until adequate answers to these two questions are forthcoming, criterion-referenced tests in reading cannot be developed.

The areas of vocabulary and word analysis can probably be handled adequately by criterion referencing, but again such tests are not generally available today.

One possible exception to the latter declarations regarding the nonavailability of good criterion-referenced tests are the Progressive Achievement Tests in vocabulary and reading comprehension published by the New Zealand Council for Educational Research and developed under the direction of Warwick Elley. These tests come very close to providing ideal answers to the questions raised above. Unfortunately, their foreign origin will probably prohibit extensive use in the United States. However, these tests deserve serious attention by test experts and reading specialists here.

Assuming that well-developed criterion-referenced tests do become available in several years, they should have much to offer in the area of performance contracting. However, to be acceptable, these tests will probably have to provide normative data to supplement the criterion method of interpretation.

The third type of test which might be selected is one which is specifically designed for the instructional materials in question. The best examples of this type of test are the end-of-unit or end-of-book

tests provided with most reading series. A variation on the end-of-unit test which may also be considered here is the continuous type of evaluation provided in computer assisted instruction (CAI) systems.

These types of tests have much to recommend them. Since they are tied closely to the instructional materials, they should have a high degree of validity for the specific program. Further, since they are spread over the entire program, they provide much greater saturation than is obtainable with the typical 30-60 minute test.

On the other hand, certain features of these tests make them unsuitable for use in performance contracts. They are quite uneven in quality. Even within a single series, quality of tests for different units and objectives varies considerably. Rarely is anything known about the reliability of such tests. Many of the questions are of the free-response type, with minimal guidance given for scoring the responses. Scorer bias becomes a major issue. Often the questions are designed to give the teacher insight about the student rather than to determine his level of learning.

Lack of normative information about end-of-unit tests is a serious problem. Thus, levels of performance on these tests are difficult to interpret. If alternate forms of the tests were available, a contract could specify that a certain raw score gain from pretest to posttest should be obtained. But I do not know of any instance where alternate forms of these tests are in fact available.

The content of the end-of-unit type tests may in fact be too closely tied to the content of the instructional materials. Sometimes exactly the same material or key phrases used in the textbook are reintroduced in the test. If "teaching for the test" is a problem when standardized tests are used, imagine the charges that would be leveled if end-of-unit tests were used for determining payment for a performance contract. More importantly, it is difficult to tell if the student can generalize beyond the specific approach used in the instructional materials (and reflected in the end-of-unit test) to other types of situations.

The latter remarks should not be construed as a general indictment of these types of tests. They are very useful for instructional purposes. Teachers, no doubt, would have to take time to construct something like these tests if they were not provided along with the instructional materials. The point being made here is that such tests do not seem to be appropriate for the type of evaluation demanded in performance contracting.

Considering kinds of tests from a somewhat different point of view, something should be said about tests of affective reactions to reading. Does the student enjoy reading? Has he developed favorable attitudes toward reading? Measurement experts have not experienced

much success in developing valid indicators of the effects of an educational program in these respects. Such indicators are clearly needed. I am not aware of any performance contract which has specified goals of an affective nature for reading. If such a contract were written, we would be hard pressed to indicate how accomplishment of the goals should be evaluated.

Levels of the Test

Most performance contracts written to date have below-average groups as target populations. What level of a test should be used with these students? If they take the level intended for typical students in their grade, scores will be quite low. This is a frustrating experience, and it adversely affects reliability of the results. On the other hand, it is often said that a lower level of a test doesn't really measure the material of interest for the grade in which the students actually are. For example, if you are working with fifth grade students, a test designed for grade three will tell you how well your students do on grade three material but not how well they do grade five material.

The latter charge may have considerable merit for subjects where curricular materials are sharply graded, e.g., arithmetic computation. However, with the exception of certain decoding skills, reading materials are not sharply graded, i.e., they do not differ greatly *in kind* from grade to grade. Of course, the materials do differ appreciably in difficulty from one grade to another. It's mainly more of the same kind of task year after year, but at successively higher levels of difficulty. Thus, in the reading area, it appears perfectly safe to use a lower level of a test for below-average pupils without seriously affecting the nature of what is being measured. The general rule to apply is to use the level of the test on which pupils will get about 50-60 percent of the items right.

Although some of the word analysis skills are graded, a judgment about which level of a test to use is generally made in terms of the appropriateness of the comprehension test. Thus, the word analysis test accompanying the comprehension test at a particular level would be used.

Norms and Other Test Information

The following observations may seem trite, but they are so frequently overlooked that they should be mentioned. The test selected should have the norms and other information called for by the contract. Or, the contract should be rewritten in terms of what the test can provide. If the contract calls for reporting growth in reading in terms of grade equivalents, then the test to be used should have

grade equivalents. Or, if the test which appears to be most valid has percentile ranks but no grade equivalents, it would be preferable to rewrite the contract so that results could be expressed in percentile ranks rather than select a less valid test which happens to have grade equivalents. Similarly, one must be sure that the test and contract are compatible in terms of other information, e.g., any expectancy data required.

One further point should be noted about norms. National norms for different tests do differ in difficulty level. Theoretically they should not, but in fact they do. This can be a serious problem for the fixed-goal type of contract. Either the contractor or the school can get "burned," depending on which series of tests is used. This is not particularly problematic if a variable goal is used; and this is precisely one of the arguments in favor of the variable goal contract.

Test Forms

For most performance contracts, it is essential that two forms of a test be available, i.e., for pretesting and posttesting. Three forms may be needed if some intermediate monitoring is to occur between pretest and posttest. For dealing with the problem of teaching for the test (see below), it would be helpful to have four or more forms.

A Final Note on Test Selection

After all of the above factors have been carefully considered, one frequently hears a question such as, "But which reading test is *really* the best?" Probably the most comforting way to answer this disarming question is to note that all of the vocabulary tests, by whatever name they go, for the major test series are highly intercorrelated, and all of the reading comprehension tests of the major series are highly intercorrelated. Thus, from the point of view of test validity, which is the most important point of view, it is difficult to make a relatively "bad" choice among the major reading tests. Various authors and research teams, working with different sets of reading materials, different item types, different points of departure, all seem to come up with measurements of the same thing.

Various tests of word analysis skills do not correlate as highly as do different tests of vocabulary or comprehension. But word analysis tests are usually not used as the main criterion of success for a reading program. They are usually used more for diagnostic purposes and evaluation of accomplishment or intermediate goals. The main criterion of success is usually the comprehension test, which is seen as measuring the final outcome, or perhaps a combination of the comprehension and vocabulary tests.

Teaching for the Test

"Teaching for the test" is one of the most vexatious matters to be treated under performance contracting. It was on precisely this issue that the Texarkana Project foundered. The spectre of this issue arises whenever accountability looms large as, for example, in state and city testing programs where results will be made public, as well as in performance contracting.

"Teaching for the test" is a paradoxical matter. If the test does not parallel the curriculum closely enough, the test is considered invalid. If the test parallels the curriculum too closely, someone will charge that teaching for the test has occurred and, therefore, the test results are invalid. This is a classical example of heads I win, tails you lose—although it is not too clear who *you* and *I* are in this instance!

It will be helpful to analyze the issue of overlap between test items and instructional materials separately for different areas of reading. The problem is most severe for vocabulary tests, particularly in the lower grades. Both tests and instructional materials depend heavily on common word lists as sources. It is little wonder, therefore, that they include many of the same words.

Elementary word analysis skills are often taught in much the same way in various reading programs because of the relative frequency of certain sound patterns in English words. For example, many reading programs teach the ending sound patterns of *-an*, *-ent*, *-ar*, *-ink*, etc. within the first year. To assure the validity of their tests, test constructors include these same basic patterns. Thus, without some very explicit evidence that a program deliberately included certain vocabulary words and certain decoding skills in relatively greater proportions because of the particular proportions used in a test, the charge that teaching for the test has occurred is difficult to justify. In fact, the charge is often nonsensical and based on ignorance of how reading programs and tests are constructed.

Tests of reading comprehension are a somewhat different matter. Because of the infinite variety of sentences and paragraphs which can be constructed on the basis of just a few hundred words, it is extremely unlikely that a test and a reading program will include identical or highly similar passages of connected discourse.

Reference has been made so far only to what is in the instructional materials for a program. What is introduced in the classroom may be a different matter. An individual teacher may review test items in an explicit or general way, thus giving the students (not to mention the contractor) an unfair advantage when the test is actually given. Obviously, this is objectionable teaching for the test. The practice, by the way, is not always engaged in with malice aforethought. I have heard some teachers say that they use the tests as workbook exercises!

There are several steps which can be taken to eliminate or, at least, alleviate this problem. The name of the test to be used could be concealed from the contractor or anyone else who might be tempted to influence the results by teaching for the test. This is not a good solution, since it is essential that the contractor and the school agree very explicitly about what would constitute a valid measure of the success of the program. However, it might be possible to reach an agreement that any one of several different tests would be acceptable. Or the test agreed upon may have several different forms. Then, if the contract is of the fixed-goal type, one of the tests or one of the forms may be selected for actual administration by some independent person without anyone else's knowledge. If the goal is of the growth type, forms or tests may be mixed for both pretest and posttest, provided that no individual gets exactly the same test on both occasions. (If tests are mixed, only group gains may be analyzed.) The latter procedures do not prevent teaching for the test. But they help to make it so time consuming (if one tries to teach for all the tests) or so risky (if one tries to guess which test will actually be used) that most people would be discouraged from trying it.

If a secure form of a test is available, use of this form helps to effectively alleviate the problem of teaching for the test. However, secure forms of major reading tests are rarely available.

In making the latter remarks, it is not assumed that the typical performance contractor is going to teach for the test. The recommendations are made for the protection of both school and contractor. Procedures which minimize the possibility of teaching for the test also minimize the possibility that accusations regarding this matter will arise. Such accusations, even if unfounded, can be as damaging to the contractor's reputation as actual proof of teaching for the test.

Special Problems in Measuring Growth

Many, perhaps most, performance contracts rely on a variable goal of the growth type. The reasons for this are varied and generally sound (Hogan, 1971). Unfortunately, the measurement of academic growth presents some very difficult problems. These problems are statistical in nature; they are not widely recognized among educators; but, if not handled adequately, they can have rather profound effects on the outcome of a performance contract. In terms of dollars, they may have a negative effect on either party to the contract, depending on specific circumstances. Fortunately, there do seem to be solutions to these problems which are neither difficult nor expensive to apply.

There are five major problems which arise in the use of tests for the measurement of academic growth. A detailed discussion of the nature of these problems, examples of them, and solutions to them are

given by Wrightstone, Hogan, and Abbott (1971). The problems are encountered so frequently in performance contracting that it seems essential to provide at least a brief description of them in this paper.

First, the performance contract with a growth goal usually specifies that growth (e.g., in reading) will equal or exceed the normal rate. "Normal" growth is almost universally defined in terms of grade equivalent (GE) units (e.g., six months gain for six months of instruction). But normal growth may also be defined in terms of percentile ranks (PR) or interval-type standard score (SS) units. Adopting either of these definitions sets quite different demands than does the GE definition. Further, the SS definition and, possibly, the PR definition, while almost never used, are theoretically more defensible than the GE definition.

The solution to this problem is a) to adopt the SS or PR definition of normal growth instead of the GE definition or b) to employ a local control group that is similar to the contract group in terms of initial status. In the latter case, peculiarities of whatever scale is used to express the amount of growth will affect the results for the control and contract groups equally.

Second, standardized tests are often used mainly because they have national norms. But most standardized tests have empirically determined norms for only one time in the school year. Other norm points are obtained by interpolation. Thus, if a performance contract calls for pretesting and posttesting within the same academic year, one or both of the sets of test results will be compared with an interpolated, not an empirically determined, norm. The interpolated norm may be higher or lower than the empirically determined norm would be. Consequently, the amount of growth obtained relative to the norm may be an over or underestimate.

To deal with this problem, it is recommended that either a) testing be done only at yearly intervals, with actual month of testing coinciding with month of standardization for the test being used, or b) one of the tests with twice-per-year standardization be used, again with test months coinciding with months of standardization, or c) a local control group be used.

Third, growth studies usually involve the use of two alternate forms and/or two adjacent levels of a test series (i.e., for the pretest and posttest). Thus, the test user must depend heavily on the equivalency of scores for different levels and forms of the test. These equivalencies are determined in prepublication research programs which are adequately conducted, and the test user can be confident about the equivalence of scores obtained from different levels and forms of a test.

If the test user has any reason to suspect the equivalence of scores,

he should employ a counterbalanced testing design in which half of the pupils receive, say, Form A as a pretest and Form B as a posttest, while the other half of the pupils receive Form B as a pretest and Form A as a posttest. Or, again, a local control group may be employed. Then any inequivalence between forms or levels will affect the results for the control group as well as for the contract group.

Fourth, growth or gain is always expressed as a difference between two scores. Differences or difference scores for individuals tend to be quite unreliable for the kinds of tests and the interest intervals usually encountered in educational situations.

The best solution to this problem is to depend on group results for the evaluation of a program. Results for the group may be expressed as an average gain or as the percentage of pupils exceeding a certain gain criterion.

Fifth, regression toward the mean may be a critical factor in some growth studies. Regression will occur when a group of students is selected as a high or low group on the basis of test results. The selected group, upon being retested, will tend to regress toward the mean of the unselected group, even though no real change has taken place. For example, a group selected to be low in reading on the basis of some reading test will tend to be somewhat higher upon being retested immediately, say, with an alternate form of the test.

When a low reading group is the target group, for example, in a performance contract, an entire group of pupils may be tested and the low readers selected for the program. Since a pretest score, which will be compared later with a posttest score, is needed, the score on the selection test is often used as the pretest score. When this is done, the difference between pretest and posttest scores will contain some admixture of real gain attributable to the program and some change due simply to the regression effect.

There are two methods of eliminating the regression effect. First, after selecting the pupils for the program, administer a separate pretest to the selected pupils. This pretest score will be compared with the posttest score. Regression will occur between selection testing and pretesting but will not affect the comparison of pretest and posttest scores. Second, certain statistical techniques may be used to estimate the expected regression which may then be subtracted from the obtained gain found between the selection test and posttest. If one wishes to employ this solution, expert statistical advice should be sought.

References

Hogan, T. P. "Some Measurement Issues in Accountability," Proceedings of the 36th Annual Conference, Educational Records Bureau. Darien, Connecticut: Educational Records Bureau, in press.

Lessinger, L. L. *Every Kid a Winner: Accountability in Education.* New York: Simon and Schuster, 1970.

Mecklenburger, J. A., and J. A. Wilson. "Performance Contracting in Cherry Creek?!" *Phi Delta Kappan,* 53 (1971), 51-58.

Millman, J. "Reporting Student Progress: A Case for a Criterion-Referenced Marking System," *Phi Delta Kappan,* 52 (1970), 226-230.

Progressive Achievement Tests. Wellington, New Zealand: New Zealand Council for Educational Research, 1969.

Tyler, R. W. "Testing for Accountability," *Nation's Schools,* 86 (1970), 37-39.

Wrightstone, J. W., T. P. Hogan, and M. A. Abbott. "Accountability in Education and Associated Measurement Problems," *Test Service Notebook No. 33.* New York: Harcourt Brace Jovanovich, 1971.

COMMENTS ON IMPERTINENT PERTINENT PROBLEMS WITH CONTENT VALIDITY IN READING TESTS:
A Response to Hogan

J. Jaap Tuinman
Indiana University

Dr. Hogan's paper deals with three somewhat disjunct topics related to their relevance to performance contracting in reading. First, Hogan discusses the nature of performance contracts in general; secondly, he analyzes some issues regarding the validity of reading tests; and, finally, he treats five major technical problems in the measurement of growth of reading achievement.

I will confine my comments largely to what I perceive to be not only the most interesting but also the most complex questions relevant to the paper—namely those subsumed under the general topic of validity problems with reading tests used in performance contracting. These questions have the additional virtue of being most central to the raison d'etre of this group of papers.

Allow me a few preliminary comments in regard to Dr. Hogan's two other concerns. I find little cause to take exception to his descriptions of either the essence of performance contracts or to their limitations. Particularly valuable seems the stress put on the complexity of real-life contracts and the necessity of continuous renegotiation. Stylistically beautiful and simple solutions to any educational problem evoke my immediate distrust.

As to Hogan's treatment of the measurement of growth problems, the reader is urged to consult Wrightstone, Hogan, and Abbott (1971). Additional discussion of similar issues can be found in excellent papers by Lennon (1971), Wardrop (1971), and Joselyn and Merwin (1972). For more technical information, Davis (1961), Harris (1963), and Cronbach and Furby (1970) should be consulted.

Contracting Amid the Great Skills Debate

So much for preliminaries. Hogan points out that goals of instruction may be differentiated in terms of curricular content. He quite correctly calls this an "exceptionally critical distinction." It is. It is so

critical that unless parties to a contract are aware of and in agreement about the particular differentiations their contract is based upon, explicitly or implicitly, there probably will be no workable contract. The necessity for content-based goal differentiation puts performance contractors in the center of what I call the great skills debate. It is their plight which I hope will provide the impetus for inching out of the stalemate at which the participants to the debate seem to have arrived. In an oversimplified but very realistic formulation, the debate focuses on two issues: the number of skills there are and whether they should be taught. Simplistic? Yes, but it is in these terms that contracts are discussed.

The Number of Skills "There Are"

There is no unique answer to the question regarding the number of skills "there are" but one must show why this is the case if one wants to aid parties to a performance contract in making up their mind about the outcomes desired. Under one interpretation of *skills* the term refers to a construct, psychologically meaningful, but by definition inferred rather than observed. Operationally such skills are frequently defined as factors in a factor analysis of correlations among performance on a variety of reading tasks. The number of this type of skills is a function of the tasks one makes the reader perform; these in turn are a function of one's definition of reading. That is not all, however. Correlations among particular tests may be unstable over time. Frequently they are a function of time spent in instruction and of the level of performance achieved. This instability may be a consequence of changes in within-group variation. Consider, for instance, the correlation between speaking vocabulary and reading vocabulary of three year olds versus that correlation for nine year olds. Then again, such instability may follow from genuine changes in integration of abilities. The determination of factors underlying performance on a variety of reading tests, therefore, is not only a function of definition but of time in the learning sequence as well.

There is a second way one can speak about skills. One finds illustrations of it in discussions about basic literacy skills. Listed as such are—more often than not—items such as ability to read traffic signs, to follow cooking directions, to fill out an application form, and so on. Skills in this sense are defined not as much by behaviors or their statistical sediments but rather by the intersection of a particular kind of reading material and, by implication, specific demands these materials make upon the reader. In order to differentiate among these skills, however, one does not customarily make these demands explicit. The question of interest is not whether or not performance on a host of such tasks could be explained by reference to a much smaller

number of factors. There is no doubt about that. Rather the question is whether the only justification for using the term *skills* is the independency among them. In my opinion, this is not the case. One obvious advantage of discourse about skills in terms of type of material read is that it allows construction of content-valid tests which have a great deal of intuitive appeal both to the test-taker and the parties using such a test as part of a contract.

The second usage of the term *skills* refers to terminal behavior, rather than to component performances. This is not the case in the third and most hotly debated use of the term. By a process of backward analysis, the reading profession has managed to come up with a great many lists of skills, which by force of logical but not necessarily psychological argument, must underly the repertoire of the accomplished reader. Lately, such lists have been expressed in behavioral terms and referred to as behavioral objectives. Such skill lists usually imply some kind of hierarchy. One frequently encounters the argument that the teaching of reading should proceed in a sequence which reflects the hierarchy suggested by the behavioral objectives lists. The *skills* to be taught are neatly specified in a most gratifying degree of concreteness. The most frequently heard argument *against* the teaching of such skills is that they have no psychological reality—that preformance on measures of these behaviors correlate so highly that their independent existence must be doubted.

Some Problems Facing Performance Contracts

Let me briefly sketch my perspective on this issue in an attempt to clarify the type of problems those engaging in performance contracting face.

1. If we consider "reading a book" a terminal behavior, there seems to be no doubt that the reader who has mastered this behavior also can show mastery of a set of less molar behaviors: turning a page, glancing from left to right, and knowing the meaning of most words.
2. En route to terminal behavior, the set of component behaviors may grow and change in that some behaviors crucial at one point may become relatively less functional later in the process.
3. The fact that someone can show mastery of a particular behavior, such as pausing after a comma when reading orally, does not reveal anything about how the behavior has been acquired. It may have been taught directly; it may have been taught indirectly; or it may be the result of a student-instigated analysis of his or another reader's behavior.

4. To demand that behaviors or skills be factorially independent in order to be considered for inclusion in the instructional process implies application of a nonrelevant criterion. If it can be shown that the teaching of skill A facilitates acquisition of skill B, justification for separate identification of A exists, even if at the moment B is acquired A and B correlate perfectly—as they would if A is a necessary and sufficient condition for B.

5. The criterion needed to make decisions about the sensibility of discriminations among skills and about the justification of hierarchies must be one which relates to the efficiency of attainment of terminal behaviors. This means that one needs empirically established hierarchies in order to refute intuitively drawn up listings of skills. It should be sufficient to refer to the work of Gagne (1968) and to Bormuth's discussion of this issue (1970) for elaboration of this comment.

6. The fact that one can establish empirical hierarchies which show that mastery of skill B is contingent upon mastery of skill A still does not imply therefore that the teaching of reading must involve the direct teaching of skills. Alternate routes, predicated on the assumption that such mastery will be indirectly acquired are logically possible. Again, however, we have neither the evidence nor very good notions about how to acquire it to permit us now to pronounce a particular detailed skills list either good or bad. Certainly, reference to factor analytic studies is of little help.

I made the six preceding points merely to illustrate that the subject matter of a performance contract in reading is open to various kinds of misinterpretations. Specification of the goals of instruction for a particular contract is a major task if it is to be anything more than an exercise in triviality. It is precisely because of this that performance contracts in reading have met with only partial success. Many measurement problems, usually cited as causes of difficulties, go back to the more basic definitional and conceptional issues hinted at above. It is, fortunately, not at all unlikely that the pressure brought on by the contract boom will result in a more general acceptance of the notion that no area of knowledge—not even one as oriented to practice as reading—can exist without proper care for the language used to describe the concepts which form its structure.

References

Bormuth, J. R. *On the Theory of Achievement Test Items.* Chicago: University of Chicago Press, 1970, 66-83.

Cronbach, L., and L. Furby. "How Should We Measure Change—or Should We?" *Psychological Bulletin,* 74 (1970), 68-80.

Davis, F. B. "The Assessment of Change," *Tenth Yearbook of the National Reading Conference.* Milwaukee, Wisconsin: National Reading Conference, 1961, 86-95.

Gagne, R. M. "Learning Hierarchies," *Educational Psychologist,* 6 (1968), 1-9.

Harris, C. W. (Ed.). *Problems in Measuring Change.* Madison: University of Wisconsin Press, 1963.

Joselyn, E. G., and J. C. Merwin. "Using Your Achievement Test Score Reports," *NCME—Measurement in Education,* 3 (1971).

Lennon, R. T. "Accountability and Performance Contracting," Invited Address at the Annual meeting of the American Educational Research Association, New York, February 1971.

Wardrop, J. L. "Some Particularly Vexing Problems in Experimentation on Reading," *Reading Research Quarterly,* 6 (1971), 329-338.

Wrightstone, J. W., T. P. Hogan, and M. A. Abbott. "Accountability and Associated Measurement Problems," *Test Service Notebook No. 27.* New York: Harcourt Brace Jovanovich, 1971.